Analysis of Phenotypic Variation in Childhood Wheezing Disorders

Graduate School for Cellular and Biomedical Sciences

University of Bern

PhD Thesis

Submitted by

Ben Daniel Spycher

from Köniz, Switzerland

Thesis advisors

PD Dr. Claudia Kuehni and Prof. Dr. Lutz Dümbgen
Institute of Social and Preventive Medicine
Medical Faculty of the University of Bern

Accepted by the Faculty of Medicine, the Faculty of Science and the Vetsuisse Faculty of the University of Bern at the request of the Graduate School for Cellular and Biomedical Sciences

Bern, Dean of the Faculty of Medicine

Bern, Dean of the Faculty of Science

Bern, Dean of the Vetsuisse Faculty Bern

Abstract

Recurrent wheezing or asthma is a common problem in children that has increased considerably in prevalence in the past few decades. The causes and underlying mechanisms are poorly understood and it is thought that a number of distinct diseases causing similar symptoms are involved. Due to the lack of a biologically founded classification system, children are classified according to their observed disease related features (symptoms, signs, measurements) into phenotypes.

The objectives of this PhD project were a) to develop tools for analysing phenotypic variation of a disease, and b) to examine phenotypic variability of wheezing among children by applying these tools to existing epidemiological data.

A combination of graphical methods (multivariate correspondence analysis) and statistical models (latent variables models) was used. In a first phase, a model for discrete variability (latent class model) was applied to data on symptoms and measurements from an epidemiological study to identify distinct phenotypes of wheezing. In a second phase, the modelling framework was expanded to include continuous variability (e.g. along a severity gradient) and combinations of discrete and continuous variability (factor models and factor mixture models). The third phase focused on validating the methods using simulation studies.

The main body of this thesis consists of 5 articles (3 published, 1 submitted and 1 to be submitted) including applications, methodological contributions and a review. The main findings and contributions were:

1) The application of a latent class model to epidemiological data (symptoms and physiological measurements) yielded plausible phenotypes of wheezing with distinguishing characteristics that have previously been used as phenotype defining characteristics.
2) A method was proposed for including responses to conditional questions (e.g. questions on severity or triggers of wheezing are asked only to children with wheeze) in multivariate modelling.

3) A panel of clinicians was set up to agree on a plausible model for wheezing diseases. The model can be used to generate datasets for testing the modelling approach.
4) A critical review of methods for defining and validating phenotypes of wheeze in children was conducted.
5) The simulation studies showed that a parsimonious parameterisation of the models is required to identify the true underlying structure of the data.

The developed approach can deal with some challenges of real-life cohort data such as variables of mixed mode (continuous and categorical), missing data and conditional questions. If carefully applied, the approach can be used to identify whether the underlying phenotypic variation is discrete (classes), continuous (factors) or a combination of these.

These methods could help improve precision of research into causes and mechanisms and contribute to the development of a new classification of wheezing disorders in children and other diseases which are difficult to classify.

Contents

List of Figures

Abbreviations

AIC	Akaike information criterion
BHR	Bronchial hyper-responsiveness
BIC	Bayesian information criterion
EM	Expectation maximisation algorithm
FM	Factor model
GWA	Genome wide association
LCM	Latent class model
MCA	Multiple correspondence analysis
RTI	Respiratory tract infection
RSV	Respiratory syncytial virus
SES	Socio-economic status

Section A: Background and Introduction

A.1. Wheezing in childhood

A.1.1 Disease burden and epidemiology

The term ‘wheeze’ refers to a whistling sound heard during breathing, indicating a modification of the airflow in the airways due to narrowing or obstruction of the airways. Together with cough and breathlessness, wheeze is a main feature of asthma. Indeed, in epidemiological surveys, asthma is often assessed by asking the participants or, in the case of young children, their parents, about their wheeze symptoms; including whether they have wheezed in the past year (designated as ‘current wheeze’) or at anytime in the past (‘wheeze ever’), or whether a physician has diagnosed asthma (‘doctor-diagnosed asthma’). Such studies find a higher prevalence for wheeze than for doctor-diagnosed asthma, particularly in children. This may be due in part to misclassification of other symptoms as wheeze. Another explanation is that there could exist, in the general population, a large and poorly recognised group of children with wheeze who do not easily fit the diagnosis of asthma and who probably rarely attend the hospital (introduction in [1]). It is well known that many children, particularly in the first years of life, suffer from wheeze only during colds and do not have other features of classic asthma such as atopy or bronchial hyper-responsiveness (BHR) [2]. In the past 25 years there has been a growing awareness that there are probably a number of different ‘wheezing disorders’ which have previously been lumped under the umbrella term ‘asthma’ [1, 3].

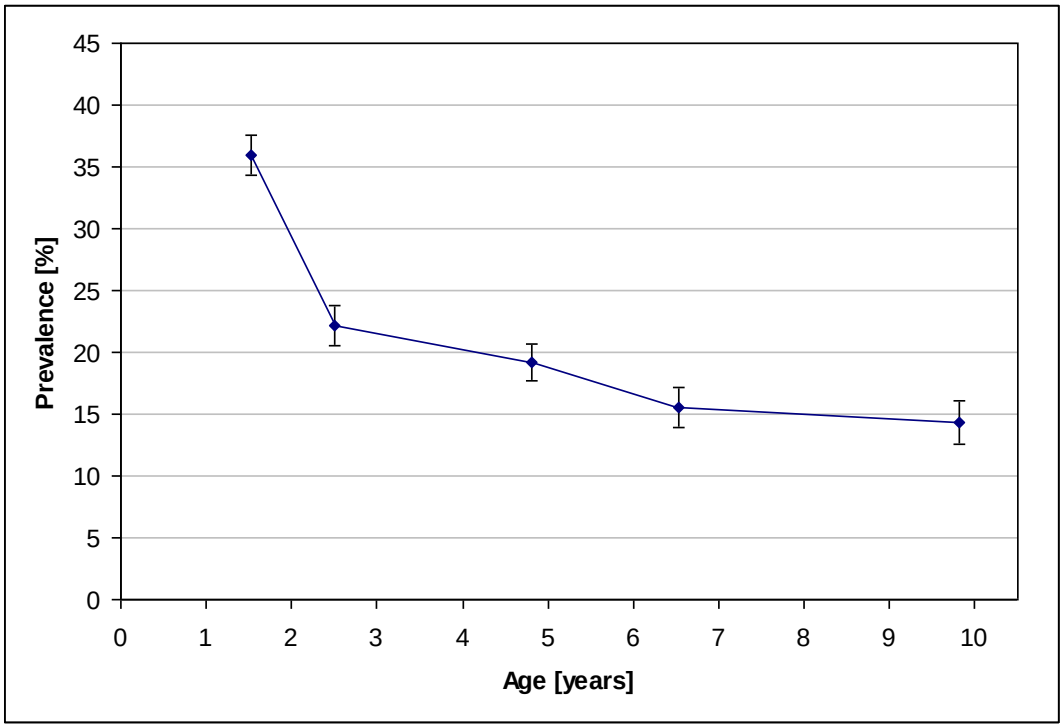
Wheezing disorders, including classic asthma, are a common problem in children. They place a considerable burden on the patients and their families, causing days missed from school and taking time away from both waged and unwaged activities [4-5]. These disorders also have a large financial impact on the healthcare system and society in general. For example, a recent study estimated total costs to the health service for 1-5 year old children with wheeze in the UK at 53 million UK pounds, representing 0.15% of the total National Health Service expenditure in the year of study, and costs to the society for caring for these children at another 2.6 million pounds [5]. Furthermore, the study included only costs incurred by children who attended hospital for wheeze or asthma, representing less than one percent of all

children in this age group. Considering that many more children are affected by wheeze but are not brought to the hospital and that children in older age groups are additionally affected by missing days from school and other activity limitations, the true impact of wheezing disorders on society is likely to be considerably larger.

The prevalence of wheezing is highest in preschool years. In a large population based cohort of children in the UK the prevalence of wheezing decreased steadily from 37% at 1 year of age to 14% at age 10 (Figure 1). In the same cohort of children, 54% reported wheeze ever by the age of 10 years. Many of the children who wheeze in the first years of life cease to have these symptoms by school-age while others continue to wheeze [6]. Patterns of onset, remission, and relapse throughout childhood (and adulthood) are highly variable [7-9]. Wheeze in childhood is associated with asthma in later life [8-10], even after prolonged remission [8-9]. Wheeze in childhood is also associated with poor lung function; which is important, considering that lung function shows considerable tracking from childhood into adulthood [11-12]. In many children wheezing in early life could therefore be a precursor of chronic respiratory diseases in later life, particularly of asthma but perhaps also of chronic obstructive pulmonary disease (COPD) [13-14].

Epidemiological studies, mainly from industrialised countries, suggest a marked increase in the prevalence of childhood wheezing during recent decades. A recent review of prevalence trends in the UK suggests that the prevalence of any recent wheeze in school-aged children increased from about 10% in the 1970s to around 20-30% at the turn of the 21st century. The increase appears to have reached a plateau in the mid 1990s and prevalence may be declining slightly since [15]. Similar rising trends reaching a plateau have been found elsewhere [16]. Kuehni and colleagues found a sharp rise in the prevalence of wheezing among pre-school children in Leicester between 1990 and 1998 (from 12% to 26% for any current wheeze) [17].

Figure 1: Prevalence of wheezing by age in a population based cohort of children (N=4300).



Data are prevalence of current wheeze (at least one attack of wheeze in the past 12 months) among respondents to 5 surveys (N is 3413, 2396, 2622, 2032, 1512 in the order of surveys) plotted against their mean age at reply. Error bars represent 95% confidence intervals.
Source: Leicester 1998 (b) cohort, which is a stratified random sample of white and south Asian children born between May 1996 and March 1997 in Leicestershire, UK [18]

The reasons for this increase are unclear. In part, it might be due to measurement problems: increased awareness among the population and changes in the diagnostic labelling may lead to more frequent usage of the labels ‘wheeze’ or ‘asthma’ and therefore to an apparent rise in the reported number of cases which is not real [16, 19]. Although trends in objectively measurable features have rarely been documented, one study has reported a parallel increase in the prevalence bronchial hyper-responsiveness (BHR) [20] suggesting that there is a real component to this rise. Changes in prevalence over such short periods of time cannot be due to changes in genetic factors, but must be due to changes in lifestyle factors and environmental exposures. A possible and frequently cited explanation, known as the ‘hygiene hypothesis’, is that increased hygiene levels in societies with a Western lifestyle

reduce the exposure to infectious agents in early life causing the immune system of children to develop an atopic response (ch. 7b in [1]). This hypothesis arose from the observation that the number of older siblings an individual had was negatively associated with atopy [21-22] and the assumption that children with older siblings are likely to have an increased and earlier exposure to infections. However, evidence suggests that less than half of the cases with wheezing in the general population are attributable to atopy [23] and trends in asthma do not consistently parallel those of atopy [20, 24] suggesting that different factors are involved in the development of asthma than in allergy in general. Using data from the Leicester cohorts, Kuehni and colleagues found that, over the period 1990 and 1998, wheezing occurring only during colds, which is typically not associated with atopy, increased at least as much in prevalence as wheezing occurring also apart from colds, which is often associated with atopy [17].

A large international multi-centre study has reported large variability in the prevalence of childhood wheezing and asthma across countries and regions, with high prevalences found mainly in English speaking countries and Latin America, a higher prevalence in Western Europe than and Eastern Europe, and relatively low prevalences in Africa and Asia [25]. An economic analysis of this data suggested that prevalence tends to be higher in high income countries than in low income countries. However, this is not a consistent pattern and there is much variability within regions. For instance, India had an average prevalence of current wheeze in 14-15 year olds of 5.8% while neighbouring Sri Lanka, also a low income country, had a prevalence of 23.0% [25]. Despite standardised methods of questionnaire translation, linguistic differences remain a major problem in such international comparisons. In many languages there is no equivalent to the English term `wheeze´, and the terms used in questionnaires may have very different connotations in the lay usage of different languages, leading to different thresholds at which people apply the terms [26-27].

A.1.2 Causes and risk factors

Wheezing is a whistling noise heard primarily during expiration. It is produced by oscillations of the bronchial wall at points of flow limitation (ch. 6a and 6b in [1]). In a

compliant tube, such as an airway, there is a maximal flow with which air can travel through which is determined by the cross-sectional area (calibre) of tube and the compliance of the tube wall. If the driving pressure increases above that required to produce flow limitation the excess pressure is dissipated through oscillations of the airway wall. In the presence of airway obstruction, this flutter may become large enough to generate audible sound, heard as wheezing (ch. 6b in [1]).

There are a large number of conditions that can cause airflow limitation leading to wheeze. Most of these are very rare at the population level. They include developmental anomalies such as bronchomalacia or host defence defects such as cystic fibrosis or ciliary diskinesia (ch. 9 in [1]). Wheeze may also be caused by localised flow limitation due to inhalation of a foreign body.

However, In most children with recurrent wheeze, the symptoms cannot be attributed to any of these specific causes. The causes of this non-specific recurrent wheezing are still poorly understood. The classical concept of wheezing suggests that wheezing is produced by airway narrowing caused by inflammatory processes, including mucosal swelling and smooth muscle activation [28]. However the degree to which these processes cause flow limitation depend on initial airway calibre and the compliance of the airway wall which in turn is a function of the elastic properties of the surrounding tissue [28].

In preschool children, wheezing is usually episodic occurring only during viral respiratory tract infections (RTI) (ch. 7b in [1]). The reasons why some children develop wheeze during viral RTI, while others do not, are still poorly understood; they may include an increased susceptibility or immune responsiveness to viral infections or poor pulmonary function (ch. 7b in [1], [28], see also article B.3 in the present thesis). Although colds are the predominant trigger of acute wheezing episodes, other factors such as laughter or crying may bring on wheeze in some children between these episodes (interval symptoms), particularly as these children get older (ch. 9 in [1], [28]). In school-aged children, wheezing is often associated with allergic sensitisation (the classic asthma phenotype) and triggered by aero-allergens (e.g. pollen, house-dust mite), however, not all children with allergy develop wheezing. Conversely, a

significant proportion of apparently non-allergic children experience wheeze episodically during viral infections only (ch. 7b in [1]).

Asthma has an important genetic component. Most genetic studies have focused on classic asthma and its related traits such as atopy and BHR, and not on the more general wheezing disorders in children. With few exceptions, twin studies report a heritability of asthma between 70% and 80% [29]. In less than two decades, genetic studies, mainly candidate gene association studies or genome-wide linkage followed by positional cloning, have identified many genes that show associations with asthma and related features such as atopy [30-31]. By 2006, 118 such genes had been identified, of which, 79 showed associations in two or more independent studies, 25 of them in more than 5 independent studies [30]. However, these reported associations are weak and those that have been replicated, were not consistently reproduced in all studies [31-34]. Since the recent advent of genome-wide association (GWA) studies, a number of new susceptibility genes have been found [35-40]. Asthma is a complex disease, i.e. it does not follow a monogenic Mendelian mode of inheritance but is rather caused by variation in numerous genes. In addition, the effects of causal genetic variants on disease-risk involve interactions with environmental stimuli and with the biological systems within the body (and, therefore, with other genes) and may occur only within limited developmental time intervals (windows of opportunity) [41-42].

A large number of factors have been studied for potential association with childhood wheezing and asthma including environmental exposures, household and lifestyle factors and person specific factors such as gender and ethnicity. Where not specifically referenced the information in the this paragraph is taken from chapter 2 in [1]. There is strong evidence that exposure to environmental tobacco smoke increases the risk of lower respiratory tract infections (RTI), recurrent wheezing and the development of asthma. Maternal smoking during pregnancy has been shown to be associated with poor lung development and diminished lung function, which in turn can predispose children to wheezing in the first years of life. Associations between air pollution, especially particulate matter from fossil fuel combustion, and respiratory symptoms have been observed in various studies. A recent study found an increased incidence of cough and wheeze in children with greater exposure to particulate pollution [43].

Another recent study also suggests that prenatal exposure to air pollution may affect lung function in newborns [44]. There is increasing evidence suggesting that exposure to allergens such as house-dust mite or cat allergen, in early life is related to the development of atopic sensitisation to that specific allergen. However, whether the allergen similarly affects the development of asthma is uncertain. A recent meta-analysis of 11 cohort studies did not find an association between asthma at school-age and pet ownership in infancy [45]. The role of socioeconomic status (SES) is context dependent. While the international comparisons suggest a greater prevalence of wheezing and asthma in affluent areas than in poor and rural areas, in the USA asthma is associated with poverty and living in the inner-city districts. However, SES is a problematic health determinant, as it is a surrogate measure for living conditions and lifestyle which may affect health outcomes in a variety of ways, such as through availability and access to healthcare, nutrition, physical exercise, housing conditions, family size, exposure to pollution and allergens. In addition, nutrition may play an important role in the development of asthma. Many studies have reported associations between breastfeeding and asthma and wheezing in childhood, however results are inconclusive, ranging from potentially protective to potentially harmful effects.

Whether or not, and by which mechanisms, infections in early life might be implicated in the development of asthma is a matter of much controversy. Various findings suggest that increased exposure to infectious agents in early childhood due to factors such as large family size [21], daycare attendance [46] or growing up on a farm [47] may protect against developing asthma. A plausible mechanism is that microbial agents may promote normal development of a T-cell system with a bias toward type 1 T-helper cells and suppression of type 2 T-helper cells which are implicated in allergen specific responses, while absence of such stimuli may allow the type 2 biased neonatal immune responses to persist and allergic sensitisation to develop (hygiene hypothesis, ch. 7b in [1]). Other observations have shown a clear association between respiratory viral infections (particularly respiratory syncytial virus (RSV)) in early life and an increased risk of asthma in later life (reviewed in ch. 7b in [1]). Two major hypotheses have been proposed to explain this association. In the first, viral infections cause the development of asthma by damaging the developing lung or affecting immune

development. In the second, the symptomatic respiratory infections are not causal but reflect an inherent susceptibility to respiratory disease including asthma. Evidence supporting either hypothesis has been reported and the matter is still much debated [48-49] (see also editorial D.4 and authors reply D.5 in the present thesis).

A.1.3 Phenotypes of childhood wheezing

Various findings related to the prevalence, natural history, physiology and risk factors of wheezing in children indicate that these conditions may not all belong to the same disease, but rather represent two or more different diseases (a review of such indications is given in [50]). In particular, wheezing occurring only during RTI ('exclusive viral-induced wheeze') predominately in young children appears to differ in many aspects from atopic asthma. In terms of the variability of airway obstruction the first is characterised by acute episodes with few or no interval symptoms while the second is characterised by a chronic obstructive component with frequent interval symptoms overlapped by acute episodes [50]. Various studies have found that wheeze occurring only in the first few years of life ('early transient wheeze') shows no statistical association with allergy, while wheezing at school-age does [6, 51]. Children with 'early transient wheeze' have been found to have diminished lung function in infancy before the onset of any lower respiratory symptoms compared to children with wheeze at school-age [6]. Various cohort studies have also shown that groups defined by association of wheeze with RTI, or time course of disease in early childhood differ in their long term prognosis [7-8, 10, 52].

The awareness that distinct diseases causing asthma-like symptoms in children might exist is not new. In the 1950s and 60s two forms of 'asthma' were commonly distinguished: 'asthmatic bronchitis' (variously called wheezy bronchitis, infant wheeze or recurrent bronchiolitis) which was characterised by wheezing and cough in association with RTIs affecting very young children, and 'asthma bronchiale' (asthma or allergic asthma) which was characterised by wheezing and shortness of breath and typically associated with allergic sensitisation [50, 53]. Around 1970, with the advent of safe and effective inhaled therapy for asthma, there was a shift toward viewing (and treating) all wheezy children as asthmatics. In the 1980s this view was replaced by a

more narrow definition of asthma which required allergy and BHR in addition to variable airway obstruction. This definition was again relaxed in the 1990s and, at the same time, the view of multiple coexisting diseases re-emerged [50]. Today, asthma is defined pragmatically, having as its main feature reversible airway obstruction causing recurrent wheeze but not requiring allergy or BHR [54]. It is regarded as a complex disease with potentially many different causes and subtypes [55] and there have been pleas to abandon the term asthma because it likely represents a heterogeneous group of diseases [3].

Because the underlying disease processes are still poorly understood, classification of wheezing disorders in children is based on phenotypes defined in terms of observable disease features (symptoms, signs and measurements). Article B.4 in this thesis reviews common phenotype definitions of childhood wheezing and methods for defining phenotypes. It reviews the two main approaches to phenotype definition: a) a uni-variate approach which selects few disease features based on expert opinion and b) a multivariate approach which applies clustering methods to observed data on multiple features to identify phenotypes in a data-driven manner.

The focus of this PhD is on the second approach of phenotype definition, i.e. on the data-driven identification of phenotypes using the combined data of many features. This approach responds to a need for more objectivity in the definition of phenotypes. The use of multivariate methods can reduce the extent of subjective choices involved in selecting the phenotype defining characteristics. The data-informed optimisation of a clustering criterion is used instead to make this selection. However, the subjectivity involved in selecting a particular clustering method and the variables to include in the analysis remains.

A.2. General aims

The broad objectives of this PhD were

a. to develop tools for analysing phenotypic variation of a disease using epidemiological cohort data, and

b. to examine phenotypic variability of wheezing among children by applying these tools to existing epidemiological data.

Some specific questions at the outset of the project were:

- Which multivariate methods are most appropriate for identifying disease phenotypes using cohort data? These data have some complicating features such as different data measurement scales (categorical , continuous, counts, and time to event data), conditional questionnaire items (e.g. questions on the pattern of symptoms asked only to subjects who have the symptom), repeated measurements and missing data.
- Are there methods that can be used to make a more objective selection of features to be included in an analysis? Many of the features are correlated and likely to represent a similar underlying dimension. How can these main dimensions be identified?
- How important are the defining criteria of traditional phenotype definition for a classification of wheeze, and how do they relate to each other? Specifically what is the role of symptom history, the main criteria used in the classification proposed by the Tucson group [6], in relation to that of triggers of wheeze, the main criteria for classification into wheeze only with, or also without, RTI [2, 50]?
- Can multivariate methods distinguish between discrete and continuous underlying variability? The former might indicate the presence of distinct disease entities while the latter might reflect variability of a single disease.

A.3. Project milestones

The articles included in the main body of this thesis (B.1-B.5) follow a chronological order reflecting the various steps of the present PhD project. These steps can broadly be grouped into three phases as listed below. The data and methods referred to here are described in sections A.4 and A.5.

1. A model for discrete phenotypes: In a first phase, we focused on the identification of distinct phenotypes of wheezing.
    a. Selection of method: From among various clustering approaches, we selected the latent class model (LCM). Reasons for this choice a listed in section A.5.1.
    b. Selection of variables to be included: We used multiple correspondence analysis (MCA, see section A.5.2) as a means of identifying important disease dimensions and to assist selection of variables for analysis (see the online supplement of article B.1 in appendix i.)
    c. Model specification and application to epidemiological data: The LCM was adapted allow for the conditional structure of questions on wheezing. Inclusion of conditional structures can be generalised to other multivariate models and is described in article B.2. We applied the model to data from the 1990 Leicester cohort (see section A.4) and results are reported in article B.1.
2. A model for discrete and continuous phenotypic variation: In a second phase, the modelling approach was extended to included continuous latent variables (see section A.5.3). This modelling framework include the LCM, the factor model (FM), and the factor mixture model (FMM) and was intended to a) allow for correlation of variables within a phenotype (e.g. due to severity gradients) and b) potentially allow distinguishing between discrete and continuous underlying variability. In several preliminary analyses this modelling framework was applied to data from the Leicester cohort 1998(a) (see section A.4). Results of these analyses are not published except in abstract form (see abstracts in appendices ii. and iii.). This modelling framework is highly flexible and models with differing structure often fit

the data similarly well. It became clear that model selection was crucial and a validation of model selection methods was needed.

3. Validating the methods: The third phase of this project focused on the question whether selection of the right model (LCM, FM, FMM) from observed data was feasible when the true model was known. This required having artificial data for which the true underlying structure was known. We decided to generate artificial data that would resemble real epidemiological data, but would originate from a known plausible model for wheezing illness in pre-school children.

   a. Obtaining a plausible model of wheezing: We set up a panel of 7 clinicians to suggest and agree on a model consisting of distinct disease entities of wheezing in preschool children. The qualitative part of this study published (article B.3).

   b. Simulation studies: Two types of simulation studies are being performed. Data are generated from three different models (LCM, FM, FMM) a) representing a) a highly controlled situation where the latent classes were at a specified and equal distance from each other (results presented in article B.5) and b) a plausible model of wheeze with general categorical data.

The developed methods are currently being applied to data from the Leicester cohort 1998(b) (section A.4). Collaboration with other cohort studies has been initiated which should allow comparison of findings across cohorts (external validation). The review article on phenotypes of childhood asthma (B.4) was written during the final year of my PhD.

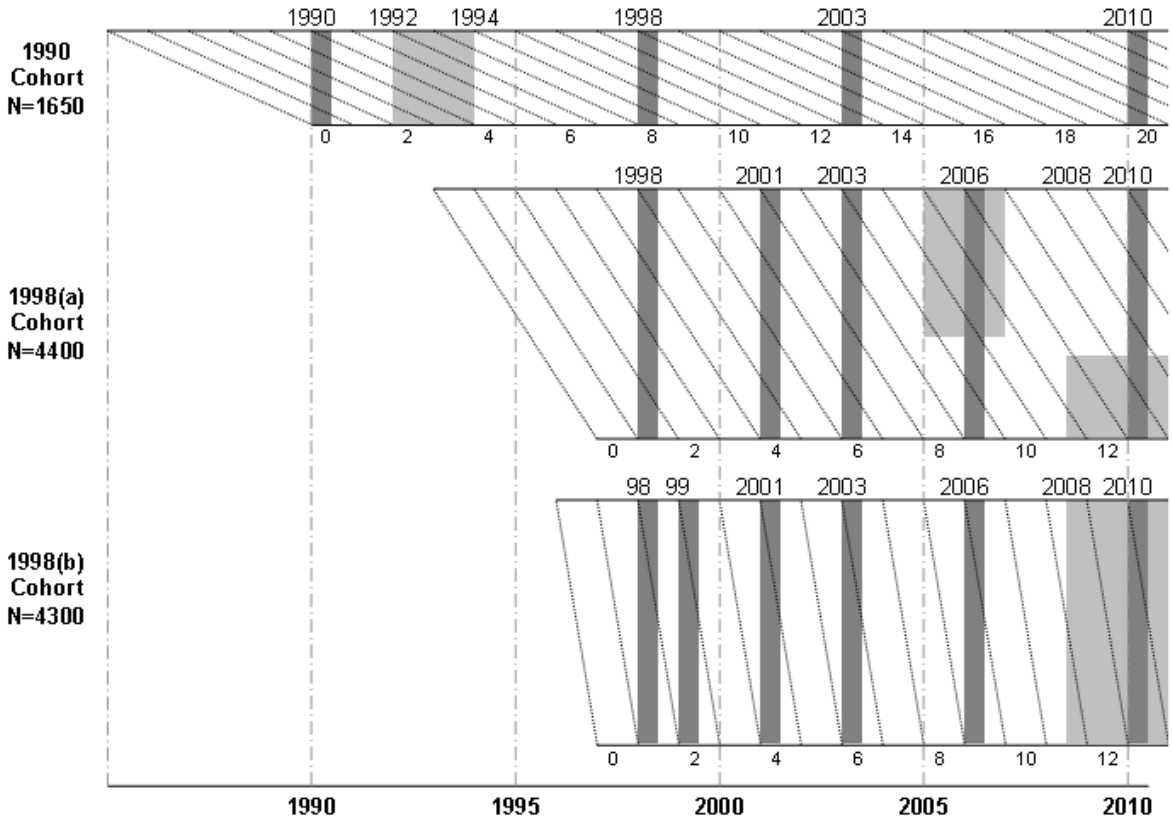
A.4. Data

The data used for this PhD project were from the Leicester respiratory cohorts. These cohorts and the collected data are described in detail in the cohort profile D.1. Briefly, the cohorts are stratified random samples from the general population born and still living in Leicester at the time of recruitment, stratified by age and ethnicity (white and south Asian). The first cohort was recruited in 1990 and consisted of 1650 white aged 0-5 years. In 1998 a large cohort was recruited consisting of 2 samples, the first (1998a) with 4400 children from a wide age span of 1-4yrs for comparability with the 1990 cohort and the second (1998b) with 4300 children all aged 1 year (Figure 2). The 1998 cohorts contained both white and south Asian children.

Data from the 1990 cohort were used for the initial study on distinguishing phenotypes (article B.1) because it provided longer follow-up (at the latest follow-up children were aged 13-18 years) which allowed comparing long-term prognosis across identified phenotypes. Data from cohort 1998(a) were used for various preliminary analyses using MCA and the latent variable modelling with both discrete and continuous latent variables (see examples in appendices ii-iv.). The methods developed are currently being applied to data from the 1998(b). This cohort has a better “resolution” for phenotype identification than the other Leicester cohorts because of more frequent follow-ups during early childhood and the a narrow age span of only 1 year which ensures age-specificity of the collected data. Currently lab measurements are becoming available for the children of this cohort.

Analysis of phenotypic variation requires a definition of the kind of data that is considered to be ‘phenotypic’. In this PhD we used as an operational definition of ‘phenotypic data’ as any data representing manifestations of the disease. This includes outcomes that are biologically linked to the disease process such as symptoms, signs and physiological measurements. However, we did not include variables that might be affected by the disease but not over a biological pathway such as treatment or doctor’s diagnosis. The full list of data collected in the Leicester respiratory studies is given in table 4 of the cohort profile D.1.

Figure 2: The Leicester Respiratory Cohorts on a time line



The X-axis represents calendar time. The horizontal bars represent the 1990 cohort and the two samples of the 1998 cohort. The (vertical) width of each bar is proportional to the number of subjects recruited. Vertical dark shaded bars represent postal questionnaire surveys. The initial survey coincides with recruitment of the cohort. Light shaded rectangles represent physiological measurements on subsamples in the respective age groups. Slanted lines demarcate age groups. The first line on the left spans the time interval during which the children were born. The numbers at the top of each cohort bar are survey years and those at the bottom child age in years.

A.5. Methods

A.5.1 Selection of methods

A large diversity of methods for the exploration of the underlying structure of data exists and various disciplines such as statistics, data mining, computer science and machine learning contribute to developing these approaches. In this PhD project, preference was given to methods based on formal statistical models. These are not inherently superior to other methods; rather, this choice reflects my inclination towards statistical methods. An advantage of statistical models, particularly relevant for this project, is that they have a natural way of handling mixed mode data, e.g. a combination of categorical, count or continuous data. Plausible statistical distributions can be specified for different types of data, for example, the multinomial distribution for categorical data or the normal distribution for appropriately transformed continuous data. Using further distributional assumptions such as (conditional) independence, the different distributions can be combined into a single model.

The aim of the first project phase was to identify distinct phenotypes of wheezing using a clustering approach. There is a large number of different clustering methods (for an extensive review see [56]). Most of these methods are based on measures of distance or proximity between data points and on an algorithm by which points of relative proximity are grouped to clusters. Given the large number of existing measures and algorithms and possible combinations thereof, the choice of any particular method can be quite arbitrary. Model-based clustering methods represent a more coherent group of methods which are based on finite mixture distributions. These models assume that the population is made up of a mixture of sub-populations specified as probability distributions.

For the first study, we decided to use a latent class model (LCM), which is a finite mixture model assuming the independence of all variables within classes (see section A.5.3). LCM is commonly, but not exclusively, used for categorical data. The advantages of using a clustering approach based on a formal model include:

- The ease of handling mixed mode data,

- The possibilities to compare models and test hypotheses, for example, to assess the appropriate number of classes (clusters), one of the major problems in cluster analysis.
- Model specification can be adapted to the particular structure of the data (e.g. longitudinal data)
- The estimation procedure can deal with missing values.

Solutions to most of these issues have also been proposed other clustering methods. For instance: the Gower distance is a measure distance for data points of mixed mode [57]; imputation methods can be used to impute missing values [58]; various cluster validation techniques have been developed to assess the appropriate number of classes [59].

In the second phase the LCM was extended to more general latent variable models which allow for continuous latent variables (factors) in addition to discrete latent classes. The problem of distinguishing between discrete and continuous latent variation, termed the problem of taxometrics, has received some attention in social and behavioural sciences [60]. Deciding whether disease varies according to distinct categories or a continuum of severity is a common problem for psychological syndromes. Several specialised methods have been developed for this purpose, but they are mainly intended for continuous data or ordinal data with many categories [60-62]. In this project, we explored model-based approaches of hypothesis testing and model selection (article B.5). Model selection using standard statistical selection criteria is increasingly being used for this purpose [63-66]. Formally testing the hypotheses of discrete vs. continuous latent variation requires tests for non-nested models. A general test for this purpose was originally proposed by Cox [67-68] and various further developments of this approach have since been made [69-70].

Throughout this work we frequently used multiple correspondence analysis (MCA) to visualise categorical data. This method was initially used to help guide variable selection and identify main disease dimensions and, in subsequent analyses, we frequently used it to visualise the discrete or continuous latent structure of both real and simulated data (an example is shown in Figure 5 in appendix iv.).

A.5.2 Multiple correspondence analysis

Multiple correspondence analysis (MCA), like principle components analysis (PCA), is a method for obtaining a low-dimensional geometric representation of high-dimensional data such that the relative position of the original data points are preserved to the largest extent possible (a unified presentation of these methods is given in [71]). While PCA is a method for continuous data, MCA is tailored to categorical data. A thorough presentation of correspondence analysis is given in [72] . The information presented here is taken mostly from chapters 4 and 5 of this reference.

Let $\mathbf{N}$ be a $s \times t$ data matrix of frequency counts. This matrix could, for example, be a contingency table, an indicator matrix[1], denoted as $\mathbf{Z}$, or a so-called Burt matrix, $\mathbf{B}$, which is the cross tabulation of an indicator matrix , $\mathbf{B} = \mathbf{Z}^T\mathbf{Z}$. Let $\mathbf{P} = \frac{1}{n}\mathbf{N}$ where n is the sum of all frequency counts in $\mathbf{N}$. Further, let $\mathbf{r} = \mathbf{P}\mathbf{1}_t$ and $\mathbf{c} = \mathbf{P}^T\mathbf{1}_s$, where $\mathbf{1}_i$ is an i-vector of ones, and let $\mathbf{R} = \mathbf{D}_r^{-1}\mathbf{P}$ and $\mathbf{C} = \mathbf{D}_c^{-1}\mathbf{P}^T$, where $\mathbf{D}_r = \mathrm{diag}(\mathbf{r})$ and $\mathbf{D}_c = \mathrm{diag}(\mathbf{c})$. The rows of $\mathbf{R}$, denoted as $\tilde{\mathbf{r}}_1,\ldots,\tilde{\mathbf{r}}_s$, are called 'row profiles' and the rows of $\mathbf{C}$, $\tilde{\mathbf{c}}_1,\ldots,\tilde{\mathbf{c}}_t$, 'column profiles'. The row and column profiles represent the rows and columns of $\mathbf{N}$ respectively, but normalised so that their elements add up to 1. The vector $\mathbf{r}$ is the centroid of the column profiles and $\mathbf{c}$ is the centroid of the row profiles. The elements of $\mathbf{r}$ represent the so-called 'masses' or the respective row profiles and the elements of $\mathbf{c}$ the 'masses' of column profiles. The point cloud for which a lower dimensional representation is sought, consists of the row profile vectors (or alternatively column profiles) in t-dimensional (s-dimensional) space with the metric $\langle\cdot,\cdot\rangle_{D_c^{-1}}$ ($\langle\cdot,\cdot\rangle_{D_r^{-1}}$), where $\langle\mathbf{a},\mathbf{b}\rangle_{D_c^{-1}} = \mathbf{a}\mathbf{D}_c^{-1}\mathbf{b}$ ($\langle\cdot,\cdot\rangle_{D_r^{-1}}$ defined analogously) for any two vectors $\mathbf{a}$ and $\mathbf{b}$ (this is a weighted Euclidean space). In analogy to the concept of inertia in physics, the inertia, denoted as I, of the point cloud is defined as the sum of the product of point masses and their squared distance from the centroid. The inertia is the same for both the column and row point clouds, i.e. $I = \sum_i r_i \left\|\tilde{r}_i - c\right\|_{D_c^{-1}} = \sum_j c_j \left\|\tilde{c}_j - r\right\|_{D_r^{-1}}$.

This can be rewriting as

[1] In an indicator matrix columns represent individual categories of the variables and rows represent individuals. Cells take on the value 1 if the respective individual has the respective category and 0 otherwise.

$$I = \text{trace}\left[\mathbf{D}_r^{-1}\left(\mathbf{P}-\mathbf{r}\mathbf{c}^T\right)\mathbf{D}_c^{-1}\left(\mathbf{P}-\mathbf{r}\mathbf{c}^T\right)\right] = \sum_i \sum_j \frac{\left(p_{ij}-r_i c_j\right)^2}{r_i c_j} = \chi^2/n\,,$$

showing that the inertia is $1/n$ times the standard χ^2-statistic associated with the data matrix $\mathbf{N}$. The objective of MCA is to find a representation of the row profiles (column) in standard Euclidean space with at most $k < \min\{s,t\}$ dimensions, such that the new row (column) points with coordinates $\tilde{\mathbf{f}}_1,\ldots,\tilde{\mathbf{f}}_s$ ($\tilde{\mathbf{g}}_1,\ldots,\tilde{\mathbf{g}}_s$), approximate the profiles in the original space in terms of their relative positions to each other. Formally, a matrix $\mathbf{X}$ is sought that minimises $\text{trace}\left[\mathbf{D}_r^{-1}\left(\mathbf{P}-\mathbf{r}\mathbf{c}^T-\mathbf{X}\right)\mathbf{D}_c^{-1}\left(\mathbf{P}-\mathbf{r}\mathbf{c}^T-\mathbf{X}\right)\right]$ among all matrices $\mathbf{X}$ of rank k or less. This can be obtained by singular value decomposition (SVD) of $\mathbf{D}_r^{-1/2}\left(\mathbf{P}-\mathbf{r}\mathbf{c}^T\right)\mathbf{D}_c^{-1/2}$. Let this SVD be $\mathbf{D}_r^{-1/2}\left(\mathbf{P}-\mathbf{r}\mathbf{c}^T\right)\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{D}_\lambda\mathbf{V}^T$ where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{D}_\lambda$ is a diagonal matrix with diagonal elements $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K \geq 0$ and $K = \min\{s,t\}$. Let $\mathbf{U}^*$ and $\mathbf{V}^*$ be the matrices with the first k columns of $\mathbf{U}$ and $\mathbf{V}$ and $\mathbf{D}_\lambda^* = \text{diag}(\lambda_1,\lambda_2,\ldots,\lambda_k)$. Then the minimising matrix is given by $\mathbf{X} = \mathbf{U}^*\mathbf{D}_\lambda^*\mathbf{V}^{*T}$ and the coordinates $\tilde{\mathbf{f}}_1,\ldots,\tilde{\mathbf{f}}_s$ and $\tilde{\mathbf{g}}_1,\ldots,\tilde{\mathbf{g}}_s$ (called 'principle coordinates') are given by the rows of $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}^*\mathbf{D}_\lambda^*$ and $\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}^*\mathbf{D}_\lambda^*$ respectively.

In the new representation, the weighted (by point masses) sum of squares of the coordinates with respect to a given axis l is $\lambda_l^{\,2}$ which can be expressed as a proportion of total inertia $I = \sum_{l=1}^{K}\lambda_l^{\,2}$. The axes are therefore ordered by the size of their contribution to "explaining" the total inertia. Some contributions to total inertia are not "interesting" because they are artefacts of redundancies in the data table $\mathbf{N}$ and the extent of such contributions differs for contingency tables, indicator or Burt matrices. Therefore adjustments are made to compute the relative contributions of each axis to the "interesting" inertia (see p. 144-5 in [72]) .

A.5.3 Latent variable modelling

Latent variable models comprise a broad group of models which emerged separately in various fields of application, particularly in the social sciences and psychology. These

'traditional' latent variable models include latent class models [73], common factor models [74-75], mixed or random effect models and various others. Their common ingredient is the inclusion of unobservable random variables, i.e. 'latent variables', in the model. Traditional latent variables models can be classified by the measurement scale (continuous or categorical) of the observed and latent variables as shown in Table 1. An extensive survey of these traditional models and their history and of modern latent variable modelling frameworks can be found in [76].

Table 1: Traditional latent variable models

	Latent variable(s)	
Observed variables	Continuous	Categorical
Continuous	Common factor model Structural equation model Linear mixed model Covariate measurement model	Latent profile model
Categorical	Latent trait model	Latent class model

Reproduced from Table 1 in [76]

The recognition that these models have a similar mathematical structure has lead to a gradual convergence of latent variable models in the past decades [76]. Unifying frameworks for latent variable modelling have been developed including the 'Generalised Linear Latent and Mixed Models' (GLLAMM), implemented in the Stata program *gllamm* [77], and the framework developed by Muthén [78-79], implemented in the program *Mplus* [80]. The model presented here and used in this project is based on the latter.

The general framework consists of two parts, a measurement part, linking the observed variables to the underlying latent variables and observed covariates, and a structural part relating the latent variables with each other and to the covariates. The observed variables are $\mathbf{y}$, a $p \times 1$ vector of continuous outcomes, $\mathbf{u}$, a $q \times 1$ vector of categorical outcomes, and $\mathbf{x}$, a $r \times 1$ vector of covariates. The latent variables are c and $\boldsymbol{\eta}$, c denoting a

categorical variable with *g* categories ('latent classes'), $c \in \{1,...,g\}$, and $\boldsymbol{\eta}$ a $d \times 1$ vector of continuous variables ('factors').

The *measurement part* defines the conditional distributions of $\mathbf{y}$ and $\mathbf{u}$ given c and $\boldsymbol{\eta}$:

i) $\mathbf{y} = \boldsymbol{\alpha}_c + \mathbf{B}_c\boldsymbol{\eta} + \boldsymbol{\Gamma}_c\mathbf{x} + \boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Theta}_c)$;

ii) the probability density (mass) functions $p(u_l \mid c, \boldsymbol{\eta}, \mathbf{x})$ of the categorical variables $u_1,...,u_q$ are modelled as logistic (binary, ordinal or multinomial) regressions on $\boldsymbol{\eta}$ and $\mathbf{x}$ with slope parameters collected in $\boldsymbol{\beta}_{ucl}$ and $\boldsymbol{\gamma}_{ucl}$ respectively and intercepts in $\boldsymbol{\alpha}_{ucl}$;

iii) conditional on the latent variables the outcomes are assumed to be independent (assumption of *local independence*), i.e. $\boldsymbol{\Theta}_c$ is diagonal matrix and

$$p(\mathbf{u}, \mathbf{y} \mid c, \boldsymbol{\eta}, \mathbf{x}) = p(\mathbf{y} \mid c, \boldsymbol{\eta}, \mathbf{x}) \prod_{l=1}^{r} p(u_l \mid c, \boldsymbol{\eta}, \mathbf{x}).$$

The *structural part* defines the probability distributions of the latent variables:

iv) $\boldsymbol{\eta} = \boldsymbol{\kappa}_c + \boldsymbol{\Lambda}_c\boldsymbol{\eta} + \mathbf{M}_c\mathbf{x} + \xi$ and $\xi \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi}_c)$;

v) $p(c \mid \mathbf{x})$ is modelled as a multinomial regression on $\mathbf{x}$ with intercepts and slopes collected in the vectors $\boldsymbol{\kappa}_0$ and $\boldsymbol{\mu}_0$ respectively.

Additionally,

vi) the random errors ξ and $\boldsymbol{\varepsilon}$ are assumed to be independent.

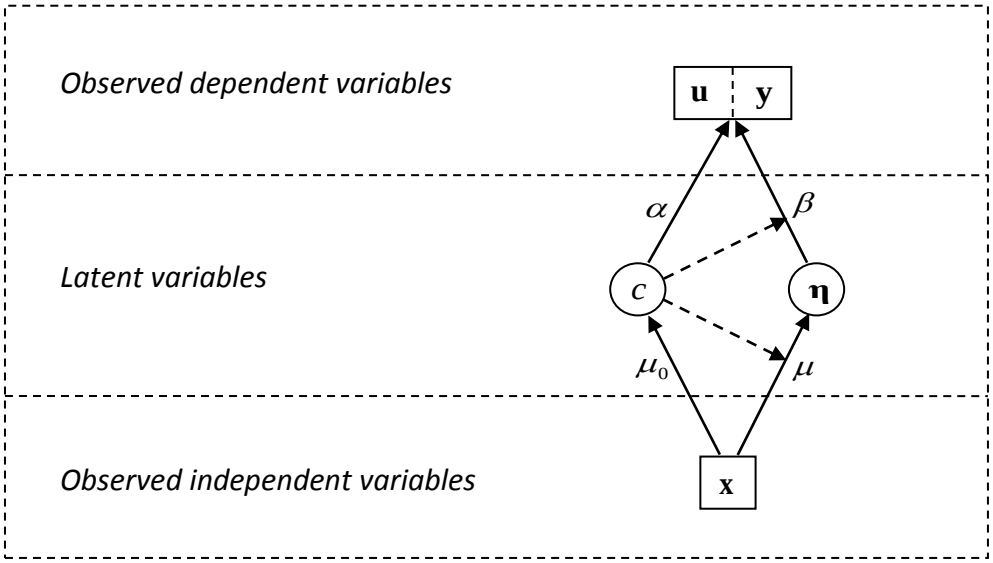
The applications in this project assume independence among latent variables. Furthermore, there are various indeterminacies in the model, e.g. the factor loadings $\mathbf{B}_c$ are not identified because pre-multiplication of the factors with a non-singular $d \times d$ matrix and post-multiplication of $\mathbf{B}_c$ with the inverse of the same matrix leaves (i) unchanged (a treatment of these indeterminacies in FMs is given [74]). Also, $\boldsymbol{\Psi}_c$ and $\mathbf{B}_c$ are not jointly identified if not further restricted. For our purposes, we simplify the structural part (iv) to

iv.a) $\boldsymbol{\eta} = \mathbf{M}_c\mathbf{x} + \xi$ and $\xi \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$,

where $\mathbf{I}$ is the identity matrix. The intercepts $\boldsymbol{\kappa}_c$ are redundant if $\boldsymbol{\eta}$ is taken out of the right hand side of (iv). There is still an indeterminacy of factor loadings due to pre-multiplication

of the factors with orthogonal matrices [74]. Typically further restrictions are placed on elements of the factors loadings $\mathbf{B}_c$ and $\boldsymbol{\beta}_{ucl}$ to reach identification.

Figure 3: Path diagram of modelling framework



Variables in boxes are observed and variables in circles are unobserved. Arrows linking variables represent regressions (linear or logistic) and are labelled with the parameters of effect size. The dotted arrows leading away from the latent class indicator *c* to other arrows indicate that the parameters of those regressions can vary over classes.

The model specified by i), ii), iii), iv.a) v), and iv) is represented in Figure 3. In the present PhD project only dependent observed variables were used, i.e. $\mathbf{u}$ and $\mathbf{y}$ (phenotypic data). However, in the future the model could include potential predictors of the latent variables such as genetic or environmental factors.

The unconditional joint distribution of the observed variables (without covariate effects) $\mathbf{u}$ and $\mathbf{y}$ is given by:

$$p(\mathbf{u},\mathbf{y}) = \sum_{c=1}^{g} \left\{ \pi_c \int_{\mathbf{R}^d} p(\mathbf{y} \mid c, \boldsymbol{\eta}) \prod_{l=1}^{r} p(u_l \mid c, \boldsymbol{\eta}) \varphi(\boldsymbol{\eta}) d\boldsymbol{\eta} \right\} \quad (1)$$

Where $\pi_c = p(c)$, $c = 1,\ldots,g$, ($p(c)$ is constant in the absence of covariates in (iv)) represents the class membership probabilities, also called 'mixing proportions', $\varphi(\boldsymbol{\eta})$ is the

probability density function of the *d*-dimensional normal distribution $\mathrm{N}(\mathbf{0},\mathbf{I})$. This is a so-called factor mixture model (FMM) [81]. If there are no continuous factors, model (1) reduces to a latent class model (LCM), given by

$$p(\mathbf{u},\mathbf{y}) = \sum_{c=1}^{g} \left\{ \pi_c \, p(\mathbf{y} \mid c) \prod_{l=1}^{r} p(u_l \mid c) \right\}. \qquad (2)$$

If there is no population heterogeneity, i.e. only one class, model (1) reduces to a factor model (FM), given by

$$p(\mathbf{u},\mathbf{y}) = \int_{\mathbf{R}^d} p(\mathbf{y} \mid \boldsymbol{\eta}) \prod_{l=1}^{r} p(u_l \mid \boldsymbol{\eta}) \varphi(\boldsymbol{\eta}) d\boldsymbol{\eta} \, . \qquad (3)$$

A major issue in specifying model (1) or (2) is choosing the appropriate number of classes. A large number of criteria has been suggested to deal with this problem (a review of such methods can be found in chapter 6 of [82]). These include standard model selection criteria such as Akaike's information criterion (AIC), or the Bayesian information criterion (BIC). Although the model with *g* classes is nested in a model with *g*+1 one classes, the likelihood ratio (LR) test does not have the usual asymptotic chi-squared distribution under the null-hypothesis of *g* classes [82-83]. A common method is to bootstrap sampling the LR statistic to obtain approximate p-values for this test .

Comparably little attention has been given to the problem of discerning more fundamental structural differences such as those existing between model (1), (2) and (3). This is the main focus of article B.5 in this thesis.

A.5.4 Model estimation and computer programs

Model (1) can be fitted by ML estimation but there are a number of issues related to the estimation of such models, some of which are briefly discussed here.

The LCM has some problem in common with the more general finite mixture models (a thorough discussion of these is given in [82]). These models are difficult to estimate as the parameters of the class distributions depend on the classes. The commonly used fitting algorithm is the expectation maximisation (EM) algorithm [84]. This algorithm was designed to fit models with incomplete data, which the case of a LCM are the unknown class

memberships. Beginning with a set of starting values for the parameters, the algorithm alternates between forming the expectation of the log-likelihood function given the observed data and current parameters by integrating out over the missing data (E-step), and maximising this expected log-likelihood with respect to the parameters (M-step). The EM algorithm generally converges to a (local) maximum. A problem is that the likelihood function of the LCM, as of the more general finite mixture models, tends to have numerous local maxima [82]. In order to find the global maximum the algorithm must be repeatedly run for different sets of starting values of the parameters, which can be selected randomly [82]. As the EM is a relatively slow algorithm this can greatly increase computation time needed for fitting a LCM. An advantage of the EM algorithm is the possibility of handling missing data [85]. This is done by integrating out over the missing data components in the E-step.

A further complication of model (1) is the presence of the combination of categorical outcomes and continuous latent variables. The M-step essentially consists of fitting a model of type (3). Without categorical outcomes the posterior distribution $p(\mathbf{y})$ has a closed form expression – in fact, it is a multivariate normal distribution. However, with categorical outcomes, there is no closed form expression and therefore numerical approximation of the integration is required; this can be computationally heavy particularly when d>1 [80].

In the first phase of this PhD project I used the *FORTRAN 77* program *MULTIMIX*[2] [86] which is designed to fit mixtures of mixed mode (continuous and categorical data). I made several extensions to this program which include the handling of missing data [87], looping over numerous repetitions of the EM algorithm with randomly sampled parameter starting values, and dealing with data arising from conditional questions (article B.2). For more general modelling in phase 2, I used the program *Mplus* (Muthén & Muthén, CA) [79-80].

I used *R* (The *R* Foundation for Statistical Computing)[3] to compute and display the multiple correspondence analyses. *R* was also frequently used for various other computations and tasks such as reading in and presenting results from latent variable modelling. I also used *Stata* (StataCorp LP, TX) for data handling and preparation and various standard analyses.

[2] Available at: http://www.stats.waikato.ac.nz/Staff/maj/multimix/ (accessed on the 5. Jan. 2010)

[3] Available at: http://www.r-project.org/ (accessed on the 5. Jan. 2010)

Section B: Publications

B.1. Article: Distinguishing phenotypes of childhood wheeze and cough using latent class analysis (*Eur Resp J* 2008; 31:974-981)

B.2. Article: Multivariate modelling of responses to conditional items: New possibilities for latent class analysis (*Stat Med* 2009;28:1927-1939)

B.3. Article: A disease model for wheezing disorders in preschool children based on clinicians' perceptions (*PLoS ONE* 2009;4:e8533)

B.4. Review article: Phenotypes of childhood asthma: are they real?

(submitted to *Clin Exp Allergy*)

B.5. Article: Distinguishing latent classes, continuous factors and their combinations with dichotomous indicators (to be submitted to *Multivariate Behav Res*)

Running head: LATENT CLASSES AND FACTORS

Distinguishing Latent Classes, Continuous Factors and Their Combinations with Dichotomous Indicators

Ben D. Spycher

Institute of Social and Preventive Medicine (ISPM)

University of Bern, Bern, Switzerland

Lutz Dümbgen

Institute of Mathematical Statistics and Actuarial Science

University of Bern, Bern, Switzerland

Ben Spycher

Institute of Social and Preventive Medicine

University of Bern

Finkenhubelweg 11

3012 Bern

Switzerland

e-mail: bspycher@ispm.unibe.ch

Abstract

Factor models and latent class models are widely used in behavioral research. Recently factor mixture models have been proposed which represent a combination of these two models. Often substantive theory cannot inform on which model is most appropriate and one would like to make this selection based on model fit to observed data. This simulation study investigates, for dichotomous outcomes, how well these models can approximate each other and thus become indistinguishable. We simulate datasets under each of the three models using different levels of separation between the classes of the latent class model. For a given level of separation, the parameters of simulation models were chosen such that the Kullback-Leibler divergence between the models was minimal. Visualization of the simulated datasets using multiple correspondence analysis showed that models approximate each other well even at high separation levels. Using standard model selection criteria, such as the AIC and BIC, models of different latent structure but similar number of parameters were well distinguishable at moderate separation levels and sample sizes. Results suggest that with categorical indicators and limited sample sizes a parsimonious parameterization of these models is crucial for their distinguishability.

Distinguishing Latent Classes, Continuous Factors and Their Combinations with Dichotomous Indicators

In the past decades, traditional latent variable models have been embedded in larger modeling frameworks (Skrondal & Rabe-Hesketh, 2007). These general models have been implemented in various software programs and such as Mplus (Muthén, 2002) or glamm (Rabe-Hesketh, Skrondal, & Pickles, 2004). Such general modeling approaches allow exploring new models which are generalizations of traditional models, for instance factor mixture models which are a combination of traditional latent class and factor models (G. H. Lubke & Múthen, 2005).

Models with discrete or with continuous latent variables follow a different modeling rationale. Continuous latent variables, i.e. factors (also called latent traits in the case of categorical indicators), are commonly used in measurement problems where it is assumed that the observed indicator variables measure an underlying gradient which is not directly observable. Discrete latent variables are commonly used in problems of typology where it is assumed that the population consists of distinct, but not directly observed, subpopulations, i.e. latent classes (also called latent profiles in the case of continuous indicators). However in many situations, which of these two models is more appropriate cannot be inferred from substantive theory and one would like this choice to be based on observed data.

The problem of differentiating between categorical and dimensional latent structure is frequently encountered in behavioral research and a number of specialized techniques have been developed to address it (see e.g. (Ruscio & Kaczetow, 2009), (Ruscio & Walters, 2009), (Meehl, 1995)). These techniques rely on detecting bimodality or discontinuities in certain transformations of the data that would indicate the existence of distinct subpopulations. They have mainly been developed for situations where the indicators are

measured on a continuous scale. Recently, Lubke and Neale have investigated the problem of distinguishing between models with different latent structures including latent class models, factor models, and factor mixture models using standard model selection criteria. They found that this distinction was unproblematic with continuous indicators (G. Lubke & Neale, 2006) but more difficult with categorical indicators (G. Lubke & Neale, 2008).

It is known that latent class and factor models can approximate each other well. For instance in the the case of continuous indicators it can been shown that a g-class model and a (g-1)-factor model are structurally equivalent with respect the model implied covariance matrix of the indicators (Molenaar & Eye, 1994). Also a single factor model is well approximated by a g class model taking on as class distributions the distributions implied by the factor model at g distinct points along the factor (Markon & Krueger, 2006). The degree to which these models can approximate each other depends heavily the separation of the latent classes or, analogously, on the impact of the factors on the indicators (factor loadings).

In this simulation study we investigated how well latent class models, factor models, and factor mixture models approximate each other with dichotomous indicators and what level of separation between latent classes is needed for them to become distinguishable using standard model selection criteria. Three levels of separation of the classes in the latent class model were assumed. For each level of the three simulation models factor models and factor mixture models were approximated to the latent class model by minimizing the Kullback-Leibler divergence between them. These models were used to generate datasets to which a set of models is fitted including the models used for data generation. The proportion of times the true model was selected over all replications was computed and compared across separation levels. We also visualize the data from the different models for a given separation level using multiple correspondence analysis.

Models

Model families

We consider following general model

$$p(x) = \text{pr}(X = x) = \sum_{i=1}^{g} \pi_i \int_{\mathbb{R}^d} \underbrace{\left\{ \prod_{l=1}^{p} \frac{\left[\exp(\lambda_{il} + \beta_{il}^T u)\right]^{x_l}}{1 + exp(\lambda_{il} + \beta_{il}^T u)} \right\}}_{=p_i(x|u)} \phi_d(u) du \tag{1}$$

where X is a p-dimensional random vector of dichotomous indicators with values 0 or 1, ϕ_d is the probability density function of the d-dimensional standard normal distribution $N_d(0, I_d)$ (I_d is the identity matrix), $d < p$, and the π_i are mixing proportions or class membership probabilities ($\sum_{i=1}^{g} \pi_i = 1$). For $d = 0$, we set $\mathbb{R}^0 = \{0\}$ and $\phi_0(0) = 1$. Model (1) can be interpreted as a latent variable model with a categorical latent variable C taking on values $i = 1, ..., g$ with probabilities $\pi_1, ..., \pi_g$ respectively and a vector of continuous factors U distributed as $N_d(0, I_d)$. Conditional on $C = i$ and $U = u$, X is distributed as $p_i(x|u)$ which assumes that the individual indicators $X_1, ..., X_p$ are independent (local independence) and models them as logistic regressions on u with intercept λ_{il} and slope β_{il}.

Based on this general model we define the the following families of models: (a) *latent class models* (LCM) with $g > 1$ and $d = 0$ (b) *factor models* (FM) with $g = 1$ and $d > 0$ (c) *factor mixture models* (FMM) with $g > 1$ and $d > 0$. Note that the factor loadings of a FMM, β_{il}, can differ over classes, that is we do not assume measurement invariance. Obviously, for LCMs the factor loadings are obsolete and are not included among the model parameters. We will use the the notation cifj to denote a model with i latent classes and j factors. Thus for instance c3f0 refers to a 3-class LCM, c1f2 to a 2-factor model and c2f1 to FMM with 2 classes and 1 factor.

Mutual approximation of model families

The extent to which the defined model families can be distinguished from each other when applied to data, depends on how well they can mutually approximate each other. If for instance the data are generated by a FMM, members of the FM or a LCM families that can approximate the true FMM closely may have a similar fit and be preferred for their parsimony.

It is clear that LCMs, FMs and FMMs are not strictly separate families, rather they overlap. For instance LCMs with up to g classes are nested in g-class FMMs - these are obtained by setting the factor loadings to zero. Similarly FMs with up to d factors are nested in d-factor FMMs - obtained by requiring factor loadings and logit intercepts λ_{il} to be equal across classes. More generally, if the models are represented by a two-dimensional array where rows and columns represent the number of factors and classes respectively, as in Figure 1, then any given model is nested in the models situated vertically downward and/or horizontally to the right of its location in the array. Furthermore, the overlap between any two models ci_1fj_1, ci_2fj_2 consists of the model at the intersection between row $j^* = \min\{j_1, j_2\}$ and column $i^* = \min\{i_1, i_2\}$, that is model ci^*fj^*. Clearly, if the data are truly from model ci^*fj^*, a pairwise comparison with a more complex model cifj, where $i > i^*$ and/or $j > j^*$, using standard model selection criteria that penalize the number of parameters should result in rejection of cifj in favor of ci^*fj^*.

In the present study we are interested in non-nested comparisons. Assume that data are generated by an unknown model which may or may not be from the general model 1. Suppose that for each of the model families LCMs, FMMs, and LCMs, we find a model that fits the data well and is not excessively parameterized. Compared to the LCM the FMM will tend to have fewer classes because it can partially compensate for this reduced flexibility by including factors. Similarly, compared to the FMM, the FM will have more factors because it needs to compensate for the complete lack of classes. In the

two-dimensional array of models, the interesting competitors for a well fitting model will be those situated downward toward the left or upward toward the right because only these can maintain model fit while keeping the number of parameters at a similar level. How well these competing models can be distinguished depends on the extent to which they can approximate each other, which essentially depends on how well discrete latent variability can be compensated through continuous latent variability and vice versa.

A LCM is a finite mixture of probability distributions identified by the distinct vectors $\lambda_1, ..., \lambda_g$ (assuming these are distinct is no loss of generality). A FM is an infinite mixture of distributions which can be indexed by the set $\{\lambda + \beta u : u \in \mathbb{R}^d\}$, where the rows of β are the β_l, $l = 1, ..., p$ (the subscript i can be dropped because there is only one class). If β is of full rank then this set is a d-dimensional affine space. Obviously, it is possible to approximate a FM with a LCM by choosing an increasing number of class distributions that populate this space. However adding classes greatly increases the number parameters of the LCM model, in our case with binary variables by $p + 1$ for each added class. Similarly a FM can approximate a LCM if sufficient factor dimensions are added such that the class distributions are included in FM's space of distributions. In our case the number of parameters of the FM increases by p for each added dimension. If the approximation is good, the choice between the two models will essentially depend on the degree with which the number of parameters is penalized.

We will measure proximity between two distributions p and q using the Kullback-Leibler divergence:

$$\begin{aligned} D_{KL}(p\|q) &= \int p(x)\log\frac{p(x)}{q(x)}dx \\ &= \int p(x)\log p(x)dx - \int p(x)\log q(x)dx \end{aligned} \tag{2}$$

This measure of divergence is convenient because maximum likelihood (ML) estimation of q corresponds to minimizing it. Suppose that P and Q are different models

and there is a distribution $q^* \in Q$ that approximates $p \in P$ by minimizing $D_{KL}(p\|q)$. Let $\hat{q}_n$ denote the ML estimator obtained by fitting model Q to a sample $\mathbf{X}_n$ of n independent draws from the distribution p. Under weak conditions, given e.g. in (Vuong, 1989), $\hat{q}_n$ exists and is a consistent estimator for q^* and $1/n$ times the likelihood ratio statistic $LR_n(p, \hat{q}_n) = \text{logL}(p|\mathbf{X}_n) - \text{logL}(\hat{q}_n|\mathbf{X}_n)$, where $\text{logL}(\cdot|\mathbf{X}_n)$ is the log-likelihood function given the data $\mathbf{X}_n$, is a consistent estimator for $D_{KL}(p\|q^*)$. Therefore $\hat{q}_n$ represents an approximation to p in terms of the Kullback-Leibler divergence.

Design of simulation study

Data Generation and Fitted Models

We generated datasets for $p = 10$ from three different models: a LCM wit 3 classes (c3f0), a FMM with 2 classes and 1 factor (c2f1) and a FM with 2 factors (c1f2) (Table 1). These models were parameterized for 3 different levels of separation between the classes of the LCM. The logit intercepts λ_i of the LCM were set such that the vectors of class specific probabilities, $p_i = \text{logit}(\lambda_i)$, were equally spaced from each other and from the vector $p_0 = (0.5, ..., 0.5)^T$ by Euclidean distance (see Appendix). This ensured that the marginal probability of each indicator was 0.5. We refer to the Euclidean distance between the p_i and p_0 as the level of separation and denote it with r. The chosen levels of separation were $r = 04$, 0.6 and 0.8. The maximum separation that can be achieved in a 4-class model while keeping the marginal probabilities at 0.5 is $r = 0.82$ (fitted models included LCMs up to 4 classes). We decided to keep within this boundary. The parameters for the FM and FMM models were chosen such that these models approximated the LCM for a given level of separation. To do this we generated a dataset of 10^5 independent draws from the LCM and fitted the FM and FMM models to this data by ML estimation. For each model and each separation level 500 datasets (4500 in all) of sample size 500 were generated.

In order to visualize the degree of separation we applied multiple correspondence analysis (Greenacre, 1984) to one of the datasets generated from each of the models. Figure 1 displays the data points for the LCM and FM at a separation of $r = 0.8$ along the two main axis. A clustering caused by the latent classes is vaguely visible in the periphery of the point cloud (left figure), while the data from the approximated FM are more evenly spread (right figure). Thus, the models are barely distinguishable by visual inspection at the largest separation level considered in this study. For lower levels the are visually indistinguishable.

To each of the generated datasets we fitted the following six models: c1f1, c1f2, c2f0, c2f1, c3f0, c4f0 (Table 1). To ensure identifiability of factor loadings, the model c1f2 was fitted with restriction of one of the factor loading parameters to zero. We did not fit more complex models such as c2f3, c2f2, or c3f1 because these required much more computation time. Table 2 shows estimated Kullback-Leibler divergence between the true models and the fitted models. To put these values into relation, consider that the Kullback-Leibler divergence between two normal distributions with variance 1 and means shifted relative to each other by θ is given by $D_{KL}(\mathrm{N}(0,1), \mathrm{N}(\theta,1)) = \theta^2/2$. A divergence of 0.02 therefore corresponds to a shift in means by $\theta = 0.2$. This is the order of divergence with which the LCM c3f0 and the FM c1f2 approximate each other at a separation level of $r = 0.6$. For a given separation level, the FMM c2f1 approximates the two other models used for data generation more closely than these approximate each other.

Model Selection

We used two standard model selection criteria (formulae in the Appendix) the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) and two variants of these previously used in a similar study, namely the sample size adjusted BIC (BICsa) and the consistent AIC (AICc) (see

(Sclove, 1987)).

We also applied Voung's LR-tests for pairwise comparisons of unnested models (Vuong, 1989). These were only computed for comparisons between the models used for data generation (which are non-nested), that is for the model pairs c3f0 vs. c1f2, c3f0 vs. c2f1 and c1f2 vs. c2f1. For a pair of models P vs. Q this test tests the following hypotheses:

$$H_0 : E\left[\log\frac{p^*(x)}{q^*(x)}\right] = 0 \tag{3}$$

$$H_p : E\left[\log\frac{p^*(x)}{q^*(x)}\right] > 0 \tag{4}$$

$$H_q : E\left[\log\frac{p^*(x)}{q^*(x)}\right] < 0 \tag{5}$$

where the expectation is with respect to the true distribution which may or may not be contained in P or Q, and $p^* \in P$ and $q^* \in Q$ are the best approximations by the Kullback-Leibler divergence to the true distribution. The null-hypothesis H_0 means that both models approximate the true model equally well and are therefore indistinguishable, while H_p (respectively H_q) means that the models p^* (q^*) is the better approximation. The test statistic is given by $1/\sqrt{n}$ times likelihood ratio $LR_n(\hat{p}_n, \hat{q}_n)$, where $\hat{p}_n$ and $\hat{q}_n$ are ML solutions, normalized by a standard error which is easy to compute. Vuong shows that under regularity conditions which are satisfied by our models, the test statistic has, asymptotically, a standard normal distribution. Values above the upper or below the lower critical value (two-sided) lead to rejection the H_0 in favor of H_p and H_q respectively. Vuong also proposed and adjusted LR-test which allows to include a penalty on the number of parameters (Vuong, 1989). We compute both the unadjusted LR-test and the adjusted LR-test using a penalty term corresponding to the BIC. We chose the 0.025 and 0.925 quantiles of the standard normal distribution as lower and upper critical values (α-level of 0.05).

The mentioned LR test assumes that the models are non-overlapping, i.e. that $P \cap Q$ is empty. As discussed above, this is not true for our models. However, for all 3 tested model pairs the overlap is not a region of interest as it neither contains the true distribution nor the best approximations p^* and q^*. It can therefore conveniently be excluded for our applications. Note that the test does not require any of the models P or Q to be the true one. In fact, because the models are non-overlapping, both are necessarily misspecified under H_0 (If either p^* or q^* is true, the Kullback-Leibler divergence between them must be zero under H_0 which can only be the case if $p^* = q^*$ meaning that the models overlap). Therefore, whenever one of the tested models is true, H_0 will tend to be rejected in favor of the true model.

Results

Latent Class Model

At the lowest separation level ($r = 0.4$), model selection criteria rarely selected the true model (c3f0) as the model of choice (Table 3). The BIC, BICsa, AICc which penalize number of parameters more heavily than the AIC tended to prefer a lower dimensional FM (f1f1) or LCM (c2f1) while the AIC frequently selected the FMM (c2f1). At medium separation ($r = 0.6$) the true model was recognized most of the times by the BIC, BICsa, and AICc. At highest separation ($r = 0.8$) the true model was almost always selected by the BIC, BICsa, and AICc, while the AIC, though recognizing the discrete class structure, was indifferent between a 3 or 4-class model. Averaged over all separation levels, the percentage of times the true model was selected was 66%, 57%, 53% and 34% for the BICsa, BIC, AICc and AIC respectively.

In pairwise comparisons using Voungs LR-test the true model (c3f0) was usually selected at medium and high separation when in competition with the FM (c1f2), but not when in competition with FMM (c2f1), which was poorly distinguishable from the true

model as indicated by frequent selection of the null-hypothesis (Table 4). When adjusting for the number of model parameters, the test was clearly in favor of the true LCM rather than the more complex FMM, but distinguishable from the FM only at the highest level of separation. When comparing the two false models, FM vs. FMM, the unadjusted test usually selected the FMM, while the adjusted test preferred the more parsimonious FM at lower levels of separation and was indifferent between the two at the highest separation.

Factor model

At the lowest level of separation the true FM (c1f2) was rarely selected by any of the selection criteria (Table 5). The BIC, BICsa, AICc tended to prefer the lower dimensional single-factor model (c1f1), while the AIC often selected the FMM. At medium separation, the true model was preferred by all criteria, particularly by the the BIC and the BICsa. At high separation, preference for the true model was unequivocal with selection probabilities above 90% for all criterion. Average percentages of selecting the true model over all separation levels were 69%, 64%, 60% and 54% for the BICsa, BIC, AICc and AIC respectively.

In pairwise testing the unadjusted LR-test was indifferent between the true FM and the LCM except at high separation where the true model was clearly preferred (Table 6). In comparison with the FMM the true model was never preverred. After adjusting the LR-test more clearly favored the true model over the LCM and was unequivically in favor of the true model when compared with the FMM. In a comparison of the false models, LCM and FMM, the FMM tended to be preferred by the unadjusted test and the LCM by the adjusted test, except at high separation where the adjusted test was indifferent between the two.

Factor mixture model

The true model (c2f1) was rarely selected at any level of separation by the BIC and the AICc which tended to prefer a single-factor model at low separation and a 3-class model at high separation (Table 7). The true model tended to be the preferred model by the AIC at all levels of separation. At high separation, the true model was selected almost 90% of the times by the AIC and was also preferred, though less clearly by the BICsa. Average percentages of selecting the true model over all separation levels were 71%, 39%, 0%, and 0% for the AIC, BICsa, BIC and AICc respectively.

In pairwise comparisons with the LCM and FM, the true FMM tended to be preferred by the unadjusted test, particularly at medium and high separation (Table 8). However, the unadjusted test almost always rejected the FMM, preferring the less heavily parameterized LCM or FM at low separation, or the null-hypothesis of indifference at high separation. In a comparison of the false models the unadjusted test tended to be indifferent at low separation and in favor of the LCM at high separation, while the adjusted test tended to be indifferent at all levels of separation.

Discussion

In summary we found that LCM and FM with dichotomous indicators were well distinguishable with standard model selection criteria except at very low separation levels of the latent classes. Conservative criteria that place a greater penalty on the number of parameters, such as the BIC and the BICsa, tended to recover these models better than the more liberal AIC. However, the FMM was poorly distinguishable from the more simple LCM and FM, particularly by the conservative criterion. Pairwise non-nested comparisons using Vuong's LR-test did not improve recognition of the true model over that attained by standard model selection criteria. The unadjusted test showed a tendency to favor the more complex FMM even when it was not the true model. The adjusted test, however was

clearly in favor of more parsimonious models LCM and FM even when these were not true.

These results demonstrate that for moderate to high levels of separation between latent classes, LCMs and their approximating FMs are easily distinguishable using standard model selection criteria. By 'moderate' separation we refer to our medium level of separation, $r = 0.6$, a level at which a visual representation of the data does not yet show a clear clustering. Only at higher levels, e.g $r = 0.8$, did the clustering become apparent (Figure 1). At the medium level of separation the FM and the LCM mutually approximated each other with a Kullback-Leibler divergence of about 0.02 which corresponds to the divergence between two standard normal distributions mean shifted by 0.2, whereas the approximation at the high separation, $r = 0.8$, corresponded to a shift in means between two normals by about 0.4. Using a standard one-sample test for differences in means assuming normality, detecting a shift in means by 0.2 and 0.4 with a power of 90% and an α-level of 0.05 would require sample sizes of approximately 260 and 70 respectively. Despite the lack of an analoguous parametric test, it might therefore be expected that latent classes of moderate separation are distinguishable from continuous factor with the sample size used in this study (n=500).

A crucial requirement for being able to distinguish latent classes from continuous factors appears to be a comparable parameterization of the competing models. In the present study the LCM c3f0 and its approximating FM c1f2 had 32 and 29 parameters to be estimated respectively. The penalties on the number parameters were therefore similar for both models, allowing the choice between the two to be based primarily on model fit. A previous study using 5-point Likert indicators found that the more conservative model selection criteria BIC, BICsa, and AICc tended to prefer FMs even when the true models were LCMs (G. Lubke & Neale, 2008). A likely reason for this is the imbalance in the number of parameter between the two models. For ordinal indicators a separate logit thresholds (for 5-point Likert indicators 4 thresholds per indicator) is needed for each

latent class, while FMs require one set of the less numerous factor loadings (1 per ordinal indicator) per included factor and only one set of thresholds. LCMs for ordinal variables with more than two categories will therefore be much more heavily parameterized than FMs of similar dimensionality.

Compared with LCMs and FMs, FMMs have the added flexibility of combining discrete and continuous latent variables. However this added flexibility may come at the cost of many additional parameters. Because the models used in the present study for data generation were all approximated to each other the FM and the LCM achieved a similar model fit as did the FMM, however with fewer parameters: 29 and 32 for the FM and LCM compared to 41 for the FMM. The additional parameters of the FMM were thus 'invested' poorly which made it difficult to recognize it as the true model, particularly by conservative criterion. We cannot, however, conclude that FMMs in general are poorly distinguishable from LCMs or FMs. Further simulation studies would be needed to investigate whether arbitrary FMMs (not approximated to the more simple structure of an LCM or FM) are distinguishable from LCMs or FMs of similar parameterization. In the present study we wanted to compare models of similar separability level when approximated by a LCM.

The present study suggests that models for categorical indicators with differing latent structures (discrete and/or continuous) can approximate each other well, such that their distinguishability crucially depends on a parsimonious parameterization. Incautious use of flexible models such as FMMs can inflate the number of parameters with only minor improvement of model fit. A careful parameterization should be sought by placing plausible restrictions on the parameters. These could e.g. include measurement invariance across classes, or setting certain factor loadings to zero. The results of this study shows that, at moderate levels of class separation and moderate sample sizes (here n=500), competing factor and class models of comparable parameterization are easily

distinguishable by standard model selection criteria.

Acknowledgments

The research of BDS was supported by Asthma UK (grant number 07/048) and the Swiss National Science Foundation (PROSPER grant 3233-069348, 3200-069349, and 823B-046481).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans Automatic Control*, *AC-19*, 716-23. (System identification and time-series analysis)

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Lubke, G., & Neale, M. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research*, *43*(4), 592-620.

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research*, *41*(4), 499-532.

Lubke, G. H., & Múthen, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21-39.

Markon, K. E., & Krueger, R. F. (2006). Information-theoretic latent distribution modeling: Distinguishing discrete and continuous latent variable models. *Psychological Methods*, *11*(3), 228-243.

Meehl, P. E. (1995). Bootstraps taxometrics - solving the classification problem in psychopathology. *American Psychologist*, *50*(4), 266-275.

Molenaar, P. C. M., & Eye, A. von. (1994). On the arbitrary nature of latent variables. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis* (p. 226-242). Thousand Oaks, CA: Sage.

Muthén, B. (2002). Beyond sem: General latent variable modeling. *Behaviormetrika*, *29*(1), 81-117.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *Gllamm manual* (Tech. Rep.). (http://www.bepress.com/ucbbiostat/paper160; accessed on 28 Dec 2009.)

Ruscio, J., & Kaczetow, W. (2009). Differentiating categories and dimensions: Evaluating

the robustness of taxometric analyses. *Multivariate Behavioral Research*, *44*(2), 259-280.

Ruscio, J., & Walters, G. D.(2009). Using comparison data to differentiate categorical and dimensional data by examining factor score distributions: Resolving the mode problem. *Psychological Assessment*, *21*(4), 578-594.

Schwarz, G.(1978). Estimating the dimension of a model. *Ann Statist*, *6*(2), 461-464.

Sclove, S. L.(1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333.

Skrondal, A., & Rabe-Hesketh, S.(2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, *34*(4), 712-745.

Vuong, Q. H.(1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*(2), 307-333.

Appendix

Parameterization of Models Used for Generating Data

Let $k = \lfloor p/g \rfloor$, i.e. the largest integer not greater than p/g (p is the number of indicators and g the number of classes). The probability vectors $p_1, ..., p_g$ of the g-LCM were set to as follows:

$$p_i = c(e_i - \bar{e}) + e_0$$

where e_i is a vector with ones in the positions $(i-1)k, (i-1)k+1, ..., ik$ and zeros otherwise, $\bar{e} = 1/m \sum_{i=1}^{g} e_i$, $e_0 = [0.5, ..., 0.5]^T$ and c is a scalar determining the separation between the p_i. This implies that the Euclidean distance between any two probability vectors p_i and p_j, $i \neq j$ has the constant value $c\sqrt{2k}$ and the distance of any p_i, $i \neq 0$, from p_0 the value $r = c\sqrt{k\frac{g-1}{g}}$.

The probability vectors were computed for separation levels r =0.4, 0.6 and 0.8. These vectors were then transformed to logits $\lambda_1, ..., \lambda_g$. For instance the parameters of

the 3-class models at separation $r = 0.4$ were as follows:

$$\lambda_1 = [0.793, 0.793, 0.793, -0.382, -0.382, -0.382, -0.382, -0.382, -0.382, 0]^T$$

$$\lambda_2 = [-0.382, -0.382, -0.382, 0.793, 0.793, 0.793, -0.382, -0.382, -0.382, 0]^T$$

$$\lambda_3 = [-0.382, -0.382, -0.382, -0.382, -0.382, -0.382, 0.793, 0.793, 0.793, 0]^T$$

Parameters for the FMM and FM were obtained by fitting these models to data generated by these LCMs as described in the main text.

Models Selection Criteria

The formulae for the four model selection criteria used are as follows:

$$AIC = -2\text{logL} + 2m$$

$$BIC = -2\text{logL} + m\text{log}(n)$$

$$BICsa = -2\text{logL} + m\text{log}\left(\frac{n+2}{24}\right)$$

$$AICc = -2\text{logL} + m\{\text{log}(n) + 1\}$$

where logL is the log-likelihood and m is the number of parameters in the model.

Author Note

Correspondence concerning this article should be addressed to Ben D. Spycher, Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, CH-3011 Bern, Switzerland E-mail:bspycher@ispm.unibe.ch)

Table 1

Design of Study: True Models Used for Sampling and Fitted Models

	Latent Classes			
Factors	1	2	3	4
0		×	☒	×
1	×	☒		
2	☒			

Note: Symbol ☐ represents true models used for generating data and × models fitted to data from each of the true models.

Table 2

Estimated Kullback Leibler Distance Between True and Fitted Models

		Fitted Model					
True model	Separation (r)	c1f1	c1f2	c2f0	c2f1	c3f0	c4f0
c3f0	0.4	0.021	0.003	0.020	0.000	0.000	0.000
	0.6	0.111	0.023	0.101	0.003	0.000	0.000
	0.8	0.347	0.092	0.310	0.017	0.000	0.000
c2f1	0.4	0.021	0.003	0.020	0.000	0.000	0.000
	0.6	0.097	0.016	0.092	0.000	0.003	0.001
	0.8	0.318	0.064	0.288	0.000	0.016	0.007
c1f2	0.4	0.017	0.000	0.017	0.000	0.001	0.000
	0.6	0.089	0.000	0.086	0.001	0.016	0.003
	0.8	0.245	0.000	0.252	0.009	0.069	0.022

Note: Data are estimated Kullback-Leibler divergences between true models and their approximation by the fitted models. Approximations are obtained by fitting the models to large random ($n = 10^5$) samples generated by the true models. Estimated Kullback-Leibler divergence is $1/n$ times the LR statistic between the true and fitted model. Negative values - which can occur in finite samples - were set to zero. These did not exceed 0.0015 in absolute value.

Table 3

Percentage of Model Choice in 500 Replications: True Model is c3f0

		Fitted Model					
Separation (r)	Criterion	c1f1	c1f2	c2f0	c2f1	c3f0	c4f0
0.4	AIC	4	7	7	50[b]	20	12
	BIC	81	0	19	0[b]	0	0
	BICsa	40	12	34	2[b]	12	0
	AICc	85	0	15	0[b]	0	0
0.6	AIC	0	1	0	37	34	28[a]
	BIC	2	22	3	0	72	0[a]
	BICsa	0	5	0	5	89	1[a]
	AICc	10	25	7	0	58	0[a]
0.8	AIC	0	0	0	3[a]	49	48[a]
	BIC	0	0	0	0[a]	100	0[a]
	BICsa	0	0	0	1[a]	98	1[a]
	AICc	0	0	0	0[a]	100	0[a]

Note: Abbreviations are as follows: r = level of separation (see Appendix); AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; BICsa = sample size adjusted BIC; AICc = Consistent AIC (formulae see Appendix). Models are denoted as cifj where i indicates the number of classes ($i = 1$ indicates a conventional factor model) and j indicates the number factors ($j = 0$ indicates a conventional latent class model). Data are the percentage (rounded to integers) of replications for which the fitted model was the first choice.

[a]model did not converge for 1 replication.

[b]model did not converge for 4 replications.

Table 4

Percentage of hypothesis choice in pairwise testing of models c3f0, c1f2 and c2f1 by Vuong's LR-Test: True Model is c3f0

	Tested Hypotheses Pair								
	c3f0 vs. c1f2			c3f0 vs. c2f1			c1f2 vs. c2f1		
Separation (r)	Hf	H0	Hg	Hf	H0	Hg	Hf	H0	Hg
Unadjusted LR-Test									
0.4	3	97	0	0	79	21	0	16	84
0.6	70	30	0	0	74	26	0	1	99
0.8	100	0	0	0	99	1	0	0	100
Adjusted LR-Test									
0.4	0	92	8	96	4	0	97	3	0
0.6	9	91	0	97	3	0	56	44	0
0.8	99	1	0	99	1	0	0	70	30

Note: For a pair of tested models F vs. G, Hf is the hypothesis that F is closer to the true model, Hg that G is closer to the true model and H0 that F and G are equally close to the true model by the Kullback-Leibler Criterion. Abbreviations and model labels are the same as in Table 3; Data are the percentage (rounded to integers) of replications for which the hypothesis was selected among the replications for which all 3 models converged (496, 500 and 499 of 500 for r=0.4, 0.6 and 0.8 respectively). Selection was based on Vuong's LR-test unadjusted and adjusted for the number of model parameters using the same penalty term as the BIC.

Table 5

Percentage of Model Choice in 500 Replications: True Model is c1f2

		Fitted Model					
Separation (r)	Criterion	c1f1	c1f2	c2f0	c2f1	c3f0	c4f0
0.4	AIC	7	11[a]	6	50[b]	16	9
	BIC	87	0[a]	13	0[b]	0	0
	BICsa	49	16[a]	28	1[b]	6	0
	AICc	89	0[a]	11	0[b]	0	0
0.6	AIC	0	60	0	33	1	5[a]
	BIC	6	93	1	0	0	0[a]
	BICsa	0	93	0	3	4	0[a]
	AICc	19	79	1	0	0	0[a]
0.8	AIC	0	91	0	8	0	1
	BIC	0	100	0	0	0	0
	BICsa	0	99	0	1	0	0
	AICc	0	100	0	0	0	0

Note: Abbreviations and model labels are the same as in Table 3. Data are the percentage (rounded to integers) of replications (500 in total) for which the fitted model was the first choice.

[a]model did not converge for 1 replication.

[b]model did not converge for 16 replications.

Table 6

Percentage of hypothesis choice in pairwise testing of models c3f0, c1f2 and c2f1 by Vuong's LR-Test: True Model is c1f2

	Tested Hypotheses Pair								
	c1f2 vs. c3f0			c1f2 vs. c2f1			c3f0 vs. c2f1		
Separation (r)	Hf	H0	Hg	Hf	H0	Hg	Hf	H0	Hg
Unadjusted LR-Test									
0.4	0	98	2	0	22	78	0	80	20
0.6	0	99	1	0	59	41	0	25	75
0.8	84	15	0	0	90	10	0	0	100
Adjusted LR-Test									
0.4	16	84	0	98	2	0	95	5	0
0.6	61	39	1	99	1	0	68	32	0
0.8	98	1	0	100	0	0	0	99	1

Note: Abbreviations and labels for models and hypotheses are the same as in Table 4; Data are the percentage (rounded to integers) of replications for which the hypothesis was selected among the replications for which all 3 models converged (483, 500 and 500 of 500 for r=0.4, 0.6 and 0.8 respectively). Selection was based on Vuong's LR-test unadjusted and adjusted for the number of model parameters using the same penalty term as the BIC.

Table 7

Percentage of Model Choice in 500 Replications: True Model is c2f1

		Fitted Model					
Separation (r)	Criterion	c1f1	c1f2	c2f0	c2f1	c3f0	c4f0
0.4	AIC	4	4	6	56[a]	18	12
	BIC	80	0	20	0[a]	0	0
	BICsa	41	8	35	3[a]	13	0
	AICc	86	0	14	0[a]	0	0
0.6	AIC	0	3	0	66	16	15
	BIC	2	53	3	0	42	0
	BICsa	0	17	0	15	67	1
	AICc	12	50	10	0	27	0
0.8	AIC	0	0	0	89	4	7
	BIC	0	3	0	1	96	0
	BICsa	0	0	0	69	29	1
	AICc	0	5	0	0	95	0

Note: Abbreviations and model labels are the same as in Table 3. Data are the percentage (rounded to integers) of replications (500 in total) for which the fitted model was the first choice.

[a]model did not converge for 10 replications.

Table 8

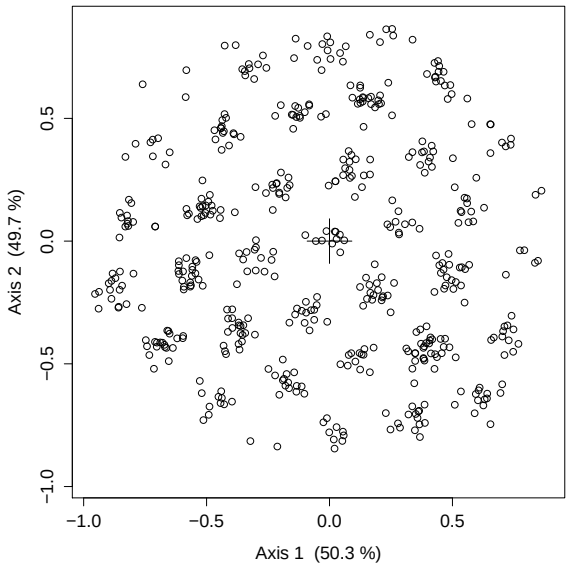
Percentage of hypothesis choice in pairwise testing of models c3f0, c1f2 and c2f1 by Vuong's LR-Test: True Model is c2f1

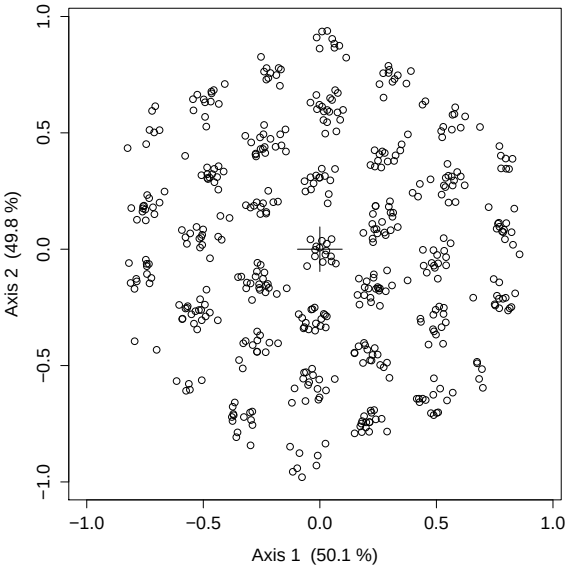
	Tested Hypotheses Pair								
	c2f1 vs. c3f0			c2f1 vs. c1f2			c3f0 vs. c1f2		
Separation (r)	Hf	H0	Hg	Hf	H0	Hg	Hf	H0	Hg
Unadjusted LR-Test									
0.4	19	81	0	85	15	0	4	96	0
0.6	55	45	0	96	4	0	32	68	0
0.8	82	18	0	100	0	0	77	23	0
Adjusted LR-Test									
0.4	0	6	94	0	5	95	0	92	8
0.6	0	9	91	0	27	73	1	97	2
0.8	0	53	47	4	95	0	38	62	0

Note: Abbreviations and labels for models and hypotheses are the same as in Table 4; Data are the percentage (rounded to integers) of replications for which the hypothesis was selected among the replications for which all 3 models converged (490, 500 and 500 of 500 for r=0.4, 0.6 and 0.8 respectively). Selection was based on Vuong's LR-test unadjusted and adjusted for the number of model parameters using the same penalty term as the BIC.

Figure Captions

Figure 1. Visual display of 500 data points generated by a latent class model (LCM) with three classes (c3f0)(left) and a factor model (FM) with 2 factors (c1f2) (right) at separation level $r = 0.8$. The data are shown by their first two principle coordinates from a multiple correspondence analysis with some added noise to avoid perfect superimposition of data points. The hexagonal shape of the point cloud and the grid-like clustering are artefacts of the discrete nature of the data (each point represents a value combination of 10 binary variables). The factor model was chosen as close as possible to the latent class model by the Kullback-Leibler divergence. The two models which are easily distinguished by standard model selection criteria are difficult to distinguish visually. The 3 latent classes are vaguely recognizable as clusters in the periphery of the point cloud of the LCM data, while the data from the FM are more evenly spread. Numbers in brackets represent the contributions of axes to the total relevant inertia. The first two axes explain almost all inertia because the latent structure of both models occupies two dimensions.


Axis 2 (49.7 %)
Axis 1 (50.3 %)
−1.0
−0.5
0.0
0.5


Axis 2 (49.8 %)
Axis 1 (50.1 %)
−1.0
−0.5
0.0
0.5
1.0

Section C: Overall Discussion & Outlook

C.1. Discussion

C.1.1 Phenotypic variation in childhood wheeze

In the first phase of this project, the focus was on identifying discrete phenotypes of childhood wheezing. We wanted to identify these in an exploratory manner using a wide range of features. We also included children with chronic cough but no wheezing. Application of LCA to data from Leicester 1990 cohort yielded 2 phenotypes of cough and 3 phenotypes of wheeze which differed in later outcomes (article B.1). The identified phenotypes of wheeze appear to confirm that dimensions which have previously been considered important for defining phenotypes, namely, the short term wheezing variability, particularly triggers of wheeze [50], and the symptom history throughout early childhood [6] are indeed relevant. The fact that two phenotypes characterised by atopic persistent wheeze and transient viral wheeze respectively were identified in data-driven manner underlines the importance of these dimensions. However, the model also yielded some unsuspected findings, in particular the finding of a persistent cough phenotype resembling an entity referred to as 'cough variant asthma'.

In a second phase, more general models (FMM) were applied that would allow for combinations of discrete (e.g. distinct phenotypes) and continuous latent variability (e. g. severity gradients). This work is ongoing and not yet published, however, preliminary results based on cross-sectional data suggest that the variability among children of wheeze related symptoms is dominated by one or two continuous disease gradients (perhaps 'symptom severity' and 'atopy'), rather than by distinct clusters (see abstract in appendix iii.). That a severity gradient does exist in observed symptom data and that it correlates with the presence of non-viral triggers is almost certain and was also repeatedly confirmed by numerous applications of MCA (see example in appendix iv.). However, whether there is also discrete latent variability needs to be investigated using more comprehensive datasets that include symptom history and physiological measurements.

These findings may appear conflicting. However, in the first study the model could not account for continuous gradients. The only possibility for the model to account for correlation between the observed variables was by fitting discrete latent classes. In the presence of a strong severity gradient it is likely that these classes would partly represent

different levels of severity. Possibly the identified phenotypes ‘atopic persistent wheeze’ and ‘transient viral wheeze’ lie at two ends of a continuous spectrum. In order to detect the presence of distinct disease entities it is therefore important that the model can accommodate severity gradients within phenotypes, e.g. using a FMM.

The findings of the present project do not provide an answer to the question of whether distinct disease entities exists within the spectrum of childhood wheezing or whether these rather represent a single disease continuum. This question is problematic as it requires a definition of disease entities (see section C.1.3). The methods used here can only address the more accessible question, which model structure best describes the observed phenotypic data. Latent classes may be indicative of distinct diseases and factors indicate the gradient of a single disease, however, they cannot provide confirmation of this.

C.1.2 Tools for studying phenotypic variation

The first phase of this project demonstrated some of the strengths of a model-based clustering approach for application to real-life epidemiological data. These include the possibility of combining data of different scales, of adapting the model to various other particular data structures (such as conditional questions) and of convenient and statistically proper [85] handling of missing data. Although having a statistical model should facilitate the choosing between different hypotheses, selecting the appropriate number of groups remains problematic. In our simulation studies (article B.5) frequently used model selection criteria were able to identify the right number of classes when the data were from a LCM and the classes were well separated. However, in a real situation, the true model is not known. It is unclear whether these criteria can detect the right number of classes when there is clustering of the data but the LCM is misspecified, i.e. is not the true model. In our application there was a large discrepancy between the number of groups chosen by the BIC (only 2 groups) and the bootstrapped LR test (5 groups) which assumes that the null-model is correctly specified. Discrepancy between the two may therefore be an indication of misspecification, but this remains to be investigated.

We proposed a method that allows including the information obtained through conditional questions without having to exclude subjects to whom the questions did not apply from the

analysis (article B.2). In our study, this allows combining detailed symptom information with long-term disease course. Subjects remain in the analysis whether or not they are symptomatic at the time points of data collection. Also, it allowed including the symptom data on wheeze together with cough symptoms without requiring all children to have wheezed. This method can be implemented using commercially available software and can be generalised to other multivariate models including the FMs and FMMs used in the second phase.

Using simulated binary data, we showed that the different latent variable models, LCM, FM, and FMM can approximate each other well, particularly when the classes are not well separated or factor loadings are modest (article B.5). At the medium level of separation considered in the simulations the FM and LCM were well distinguishable using the BIC or the sample size adjusted BIC. The FMM was more difficult to distinguish because of the greater number of parameters and the fact that the particular model we used could well be approximated by a LCM with fewer parameters. This demonstrates the importance of parsimonious parameterisation when comparing such models. We have not yet tested model selection procedures using data from a plausible model of wheezing diseases (article B.3). Judging from these first simulation results, the ability to identify disease entities will depend largely on the degree of separation between the corresponding phenotypes. FMMs should be used with caution as the number of model parameters increases rapidly with the number of classes. Plausible restrictions on the parameters should be implemented, such as setting factor loadings to zero on variables that are not thought to be important indicators of underlying gradients.

This project also showed some limitations of latent variable modelling. Latent variables models can be time consuming to fit. Models with many classes require many repeated runs of the EM algorithm using different starting values due to numerous local maxima. Factor models with categorical outcomes require numerical integration for evaluation of the likelihood function. The combination of factors and classes in the FMM can be particularly heavy on computation. For fitting a single model to one data set this is not a major limitation, however, if bootstrapping is required, computation times may be prohibitive. Running the bootstrap LR tests for the LCM in initial study required several days. As all parametric models, latent variable models require specification. Assumptions are made

regarding the distribution of the variables, such as the assumption of local independence. These assumptions may be unrealistic and restrictive. However, they are transparent and can, in some cases, be tested. Also, having a parametric model allows incorporating prior information about the data into the model (such as the conditional questions). We did not explore other methods for exploring the underlying structure of data (such as those reviewed in [56]) and cannot say how these would compare with latent variable modelling. What can be said for all these exploratory methods, is that, although more objective than expert opinion, they are not entirely objective. All stages require subjective decisions, such as the selection of variables and appropriate variable transformations, the choice of the methodological approach, of the optimisation criteria and solving algorithm.

Throughout this project MCA was used to visualise data. One tends to have greater confidence in the existence of structure in the data if this structure can perceived by one's own senses. Categorical (nominal and ordinal) data are particularly difficult to visualise as they lack a unit of measurement. Although the categories of a variable can be assigned to distinct points on the real axis, the positions of these points are completely arbitrary for nominal data, and arbitrary up to their order for ordinal data. The shape of a point cloud of multivariate categorical data would not be meaningful if such coordinates were used. In MCA, more meaningful coordinates are computed such that the point cloud reflects the essential associations in the data matrix. In this project, MCA was useful for variable selection prior to modelling because it allowed identifying groups of similar categories from different variables that might be well represented by the categories of just one variable. It also revealed structures such as gradients of severity (see Figure 5 in appendix iv.). However, in many instances these structures were not easily recognisable in projections on the 2-dimensional plane and the method frequently recovered artifactual structure. For instance, if children tended to have missing responses to similar questions and these were coded with a 'missing' category, this would create artificial associations which affected the shape and orientation of the point cloud. There are some remedies to this problem such as the reweighting 'missing' categories or imputing missing values [72]. Also, we found that MCA is of limited use for distinguishing between discrete and continuous latent variation as model selection criteria tend to detect these differences earlier (article B.5)

C.1.3 Validation of phenotypes

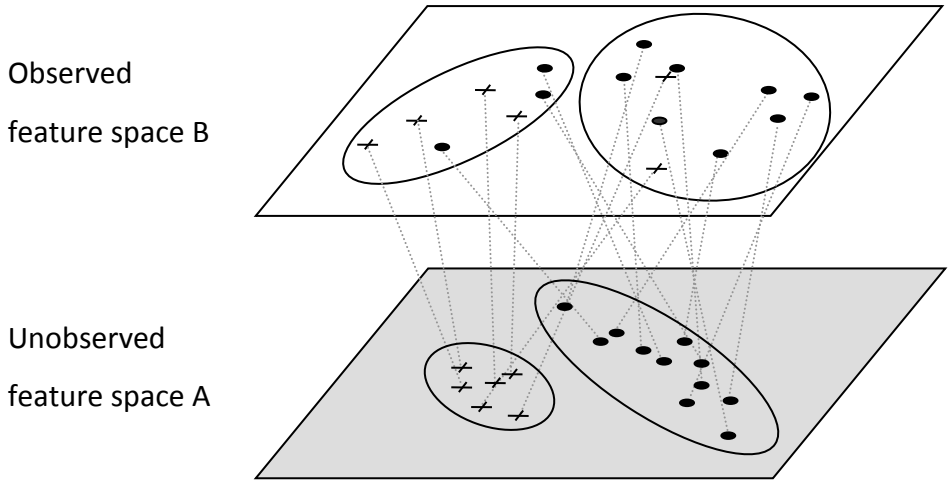
If phenotypes are assumed to be real they require validation. The term phenotype is used in different ways in the literature on asthma and childhood wheeze: While it is sometimes simply used as a synonym for an observable feature, it is often used to represent a hypothesized disease entity (review article B.4). The criteria by which one could verify whether a phenotype does represent a real disease entity are unclear. A strong criterion for considering a disease as a separate entity is, if it has a distinct aetiology. This is, for example, spelled out in the following citations: “Classification of disease usually proceeds from the general to the specific, from the phenotypic classification […] to a pathogenetic classification […] and eventually to an aetiologic classification.” (p. 56 in [88]) and “[…] the characteristics specifying the population of interest may be an etiologic agent, a specified disorder of structure and function, or a consistent syndrome. These four levels indicate progressively decreasing knowledge of the disease and therefore decreasing priority as defining characteristics: aetiology has highest priority, altered structure or function, respectively, have intermediate priority, and clinical features have the lowest priority.” (p. 679 in [89]).

The aetiology and pathophysiology of childhood wheezing or asthma are poorly understood and therefore disease classification occurs on the phenotypic level. Various criteria have been used to justify proposed phenotypic classifications, such as whether the phenotypes are associated with other clinical features not used to define them, with long term prognosis or with distinctive risk factors (review article B.4).

The implicit assumption of such ‘validation’ attempts is that certain phenotypes are, in some way, more ‘true’ than others. What could be meant by ‘true phenotypes’ is illustrated in Figure 4. The white plane (B) represents the feature space measured in a particular study, i.e. all the disease related features on which data was collected, and each of the points (circles and crosses) represents the particular combination of features measured in a subject. To state that there are true ‘true phenotypes’ implies the existence of some feature space (A) that, if measured, would reveal two obviously distinct groups indicated by the ovals in (A) (circles and crosses are used to represent the true classification). We would call these entities ‘true’ because everybody would agree to their distinction if shown the feature space (A). Because the true structure is not perfectly mapped into the observed space the clustering is less distinct in (B). Based on the clusters observed in (B) the subjects might be

grouped into two phenotypes as indicated by the ovals. Although there is some misclassification, the proposed phenotypes reflect the true classification to some extent. The proposed phenotypes are in this sense 'more true' than other arbitrarily defined groupings.

Figure 4: Schematic illustration of true phenotypes



The observed feature space (B) consists of all the disease related features (e.g. symptoms and measurements) on which data are collected in a particular study. The two clusters observed in (B) could be used to define two phenotypes (ovals). The statement that these are 'true phenotypes' suggests that there is some feature space (A) in which this separation would be obvious if that space were observed (circles and crosses represent the true classification). Feature space (A) might consists of disease markers revealing two distinct biological pathways. In the observed feature space (B) this distinction is blurred. If among all feature spaces, there is no self-evident separation as in (A), then the existence of phenotypes and their definitions can always be disputed.

Validation, in a strong sense, would mean to find and measure variables with respect to which the distinction between phenotypes is self-evident. This can be thought of as finding a space (A). The features of (A) could be the particular environmental and genetic factors at disease inception or at the time point when the different entities begin to diverge combined with the phenotypic features revealing the different disease course. If these features are found these would become the disease defining characteristics.

Validation, in a weak sense, would mean to find indications for the existence of a feature space (A). The feature space (A) may be unknown or difficult to measure. However one can attempt to measure other spaces (different from (B)), e.g. long term outcomes, or response to treatment, to see if in these new spaces the subjects are similarly clustered. Or one could sample new subjects in the same feature space (B), i.e. make the same measurements in a different sample, to see whether there is again a clustering suggestive of the same phenotypes. A clustering into similar phenotypes repeatedly observed in different spaces or in different samples is not likely to happen by chance. If the causes for this separation into phenotypes were known and the different pathways observed a clear picture as in (A) would emerge.

In the first study we compared differences in prognosis among children at later follow-up (article B.1). We have not yet performed any external validation studies using different cohorts. Both of these approaches would be examples of weak validation. A more promising approach may be to search for new feature spaces that may reveal clearer structure (approaching features space (A)). This would mean to identify markers of the distinguishing mechanisms and pathways.

C.2. Outlook

C.2.1 Identifying phenotypes of wheeze

This project is still ongoing. Several further steps are planned or underway:

- Simulation studies using artificial data from a plausible model of childhood wheeze (article B.3)
- Application of modelling approach to data from the Leicester cohort 1998(b). This cohort has a better resolution of symptom data and physiological measurements are becoming available for large subsamples of children. This will also allow external validation of the phenotypes identified in the first study using data from the 1990 cohort
- Validation of findings in other European cohorts. The Leicester cohorts study is participating in the Global Allergy and Asthma European network (GA2LEN) initiative on cohort studies [90-91]. A main aim of this initiative is to combine research activities between all European birth cohorts on asthma and allergy. It therefore provides an excellent platform for comparison of findings across different cohorts.

There is a need for new cohort studies that contribute new information relevant for understanding disease heterogeneity. Exploratory methods such as the ones used in this project are limited by the available data. If the measured features poorly distinguish between different underlying disease entities the latter will be difficult to detect, i.e. there will be a strong overlap between phenotypes. Therefore, markers need to be identified that would distinguish between different mechanisms and pathways. This can follow either a) a hypothesis-free (or black-box) approach in which a large number of signals are tested for association with disease (e.g. genome wide association studies, exhaled volatile organic compounds [92]) or b) a hypothesis-driven approach (article B.3) in which measurements targeting specific hypothesised mechanisms (e.g. candidate gene association studies, exhaled nitric oxide [93]).

C.2.2 Genetic association studies

Improved phenotype definitions could help improve precision in genetic studies. Candidate gene studies and, in recent years, genome wide association studies (GWAS) have identified a number of new genes that appear to be implicated in the development of asthma [30-31, 35-40]. However the measured effect sizes are small and many findings could not be replicated [31-34]. Phenotype definition is one among a number of factors that complicate the discovery of causal genetic variants [41, 94]. If phenotypes used in genetic studies well reflect the main underlying disease processes they will be more specific for the effects of causative variants. To correct for multiple testing, GWAS typically adopt low significance levels for individual tests [95]. Detecting effects at these levels requires large sample sizes. If the measured effects are diluted due to poor definition of the outcome the requirements on sample size and costs are greatly increased. In such a situation, investing resources into improved measurement of phenotypes may be more cost-effective than increasing sample size.

Latent variable modelling may be a useful method for defining phenotypes in large studies that measure numerous phenotypic features. Instead of selecting a single feature as outcome or subjectively defining an outcome based on a combination of features, the estimated values of the latent variables (factor scores for factors, and class membership for latent classes) of an appropriate model could be used. In a well specified model, these variables explain the associations between the measured features and may therefore better reflect the underlying disease processes that give rise to these features. Model selection methods (article B.5) could help determining the appropriate latent structure of the model. If a FM is more appropriate than a LCM, factor scores should be used instead of discrete phenotypes. Classifying subjects into discrete phenotypes in this situation could greatly reduce precision.

Another approach would be to extend the modelling framework to include covariate effects (see Figure 3). Thus genetic (only a selection of variants could be included) or environmental factors could be included in the model. The latent classes of factors would then represent mediating variables between risk factors and outcomes. These might reflect actual disease processes.

C.2.3 Clinical relevance

In clinical practice there is a strong demand for a reliable phenotypic classification, particularly for preschool children. As the panel study carried out in this project showed, there is wide agreement among clinicians that wheezing disorders in childhood comprise different disease entities, but a range of different concepts for classifying them exists (article B.3). A popular phenotypic classification distinguishes between episodic (viral) wheeze occurring only during colds and multiple trigger wheeze (as proposed in recent treatment recommendations [96]). It is possible that these phenotypes do in fact reflect different underlying diseases [50]. However, it is unlikely that there is an exact correspondence between underlying diseases and this uni-dimensional phenotypic definition.

Exploratory analysis of multivariate phenotypic data could help identify important other features to be included in the clinical definitions. As discussed in the review article (B.4) including more features in a definition can be problematic as the definition may become too exclusive. Softer diagnostic tools, perhaps based on a probabilistic classification or on decision trees might be a solution. However, the clinical benefits (response to treatment, predicting long term outcome) of such tools would have to be shown.

Ultimately, classification of wheezing in childhood needs to move from the phenotypic level to a pathophysiological and aetiological level. The approach developed in this project can contribute to this process by improving precision in studies on mechanisms and causes of disease. As the picture becomes clearer, phenotypic labels are likely to be replaced by a diagnosis based on specific markers of the underlying disease processes.

References

1. Silverman M, Childhood Asthma and Other Wheezing Disorders. London: Arnold, 2002.
2. Silverman M, Out of the mouths of babes and sucklings: lessons from early childhood asthma. Thorax 1993;48: 1200-1204.
3. A plea to abandon asthma as a disease concept. Lancet 2006;368: 705.
4. Taylor WR, Newacheck PW, Impact of childhood asthma on health. Pediatrics 1992;90: 657-662.
5. Stevens CA, Turner D, Kuehni CE, Couriel JM, Silverman M, The economic impact of preschool asthma and wheeze. Eur Respir J 2003;21: 1000-1006.
6. Martinez FD, Wright AL, Taussig LM, Holberg CJ, Halonen M, Morgan WJ, Asthma and wheezing in the first six years of life. N Engl J Med 1995;332: 133-138.
7. Morgan WJ, Stern DA, Sherrill DL, et al., Outcome of asthma and wheezing in the first six years of life: follow-up through adolescence. Am J Respir Crit Care Med 2005;172: 1253-1258.
8. Stern DA, Morgan WJ, Halonen M, Wright AL, Martinez FD, Wheezing and bronchial hyper-responsiveness in early childhood as predictors of newly diagnosed asthma in early adulthood: a longitudinal birth-cohort study. Lancet 2008;372: 1058-1064.
9. Strachan DP, Butland BK, Anderson HR, Incidence and prognosis of asthma and wheezing illness from early childhood to age 33 in a national British cohort. BMJ 1996;312: 1195-1199.
10. Edwards CA, Osman LM, Godden DJ, Douglas JG, Wheezy bronchitis in childhood: a distinct clinical entity with lifelong significance? Chest 2003;124: 18-24.
11. Stern D, Morgan W, Wright A, Guerra S, Martinez F, Poor airway function in early infancy and lung function by age 22 years: a non-selective longitudinal cohort study. Lancet 2007;370: 758-764.
12. Phelan PD, Robertson CF, Olinsky A, The Melbourne Asthma Study: 1964-1999. J Allergy Clin Immunol 2002;109: 189-194.
13. Silverman M, Kuehni CE, Early lung development and COPD. Lancet 2007;370: 717-719.
14. Bush A, Asthma research: the real action is in children. Paediatr Respir Rev 2005;6: 101-110.
15. Anderson HR, Gupta R, Strachan DP, Limb ES, 50 years of asthma: UK trends from 1955 to 2004. Thorax 2007;62: 85-90.
16. Akinbami LJ, Schoendorf KC, Trends in childhood asthma: prevalence, health care utilization, and mortality. Pediatrics 2002;110: 315-322.
17. Kuehni CE, Davis A, Brooke AM, Silverman M, Are all wheezing disorders in very young (preschool) children increasing in prevalence? Lancet 2001;357: 1821-1825.
18. Kuehni CE, Brooke AM, Strippoli MF, Spycher BD, Davis A, Silverman M, Cohort profile: the Leicester respiratory cohorts. Int J Epidemiol 2007;36: 977-985.
19. Magnus P, Jaakkola JJ, Secular trend in the occurrence of asthma among children and young adults: critical appraisal of repeated cross sectional surveys. BMJ 1997;314: 1795-1799.
20. Peat JK, van den Berg RH, Green WF, Mellis CM, Leeder SR, Woolcock AJ, Changing prevalence of asthma in Australian children. BMJ 1994;308: 1591-1596.
21. Strachan DP, Hay fever, hygiene, and household size. BMJ 1989;299: 1259-1260.

22. von Mutius E, Martinez FD, Fritzsch C, Nicolai T, Reitmeir P, Thiemann HH, Skin test reactivity and number of siblings. Bmj 1994;308: 692-695.

23. Pearce N, Pekkanen J, Beasley R, How much asthma is really attributable to atopy? Thorax 1999;54: 268-272.

24. von Mutius E, Weiland SK, Fritzsch C, Duhme H, Keil U, Increasing prevalence of hay fever and atopy among children in Leipzig, East Germany [see comments]. Lancet 1998;351: 862-866.

25. Lai CK, Beasley R, Crane J, Foliaki S, Shah J, Weiland S, Global variation in the prevalence and severity of asthma symptoms: phase three of the International Study of Asthma and Allergies in Childhood (ISAAC). Thorax 2009;64: 476-483.

26. Netuveli G, Hurwitz B, Sheikh A, Lineages of language and the diagnosis of asthma. J R Soc Med 2007;100: 19-24.

27. Crane J, Mallol J, Beasley R, Stewart A, Asher MI, Agreement between written and video questions for comparing asthma symptoms in ISAAC. Eur Respir J 2003;21: 455-461.

28. Frey U, Why are infants prone to wheeze? Physiological aspects of wheezing disorders in infants. Swiss Med Wkly 2001;131: 400-406.

29. Los H, Postmus PE, Boomsma DI, Asthma genetics and intermediate phenotypes: a review from twin studies. Twin Res 2001;4: 81-93.

30. Ober C, Hoffjan S, Asthma genetics 2006: the long and winding road to gene discovery. Genes Immun 2006;7: 95-100.

31. Vercelli D, Discovering susceptibility genes for asthma and allergy. Nat Rev Immunol 2008;8: 169-182.

32. Rogers AJ, Raby BA, Lasky-Su JA, et al., Assessing the reproducibility of asthma candidate gene associations using genome-wide data. Am J Respir Crit Care Med 2009;179: 1084-1090.

33. Daley D, Lemire M, Akhabir L, et al., Analyses of associations with asthma in four asthma population samples from Canada and Australia. Hum Genet 2009;125: 445-459.

34. Willis-Owen SA, Cookson WO, Moffatt MF, Genome-wide association studies in the genetics of asthma. Curr Allergy Asthma Rep 2009;9: 3-9.

35. Moffatt MF, Kabesch M, Liang L, et al., Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 2007;448: 470-473.

36. Hui J, Oka A, James A, et al., A genome-wide association scan for asthma in a general Australian population. Hum Genet 2008;123: 297-306.

37. Ober C, Tan Z, Sun Y, et al., Effect of variation in CHI3L1 on serum YKL-40 level, risk of asthma, and lung function. N Engl J Med 2008;358: 1682-1691.

38. Choudhry S, Taub M, Mei R, et al., Genome-wide screen for asthma in Puerto Ricans: evidence for association with 5q23 region. Hum Genet 2008;123: 455-468.

39. Imada Y, Fujimoto M, Hirata K, et al., Large scale genotyping study for asthma in the Japanese population. BMC Res Notes 2009;2: 54.

40. Himes BE, Hunninghake GM, Baurley JW, et al., Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene. Am J Hum Genet 2009;84: 581-593.

41. Guerra S, Martinez FD, Asthma genetics: from linear to multifactorial approaches. Annu Rev Med 2008;59: 327-341.

42. Martinez FD, Genes, environments, development and asthma: a reappraisal. Eur Respir J 2007;29: 179-184.

43. Pierse N, Rushton L, Harris RS, Kuehni CE, Silverman M, Grigg J, Locally generated particulate pollution and respiratory symptoms in young children. Thorax 2006;61: 216-220.

44. Latzin P, Roosli M, Huss A, Kuehni CE, Frey U, Air pollution during pregnancy and lung function in newborns: a birth cohort study. Eur Respir J 2009;33: 594-603.

45. Carlsen KCL, Roll S, Carlsen KH, et al., Pets in infancy - asthma or allergy at school age? Collaborative meta-analysis of individual participant data from 11 European birth cohorts. submitted 2009.

46. Ball TM, Castro-Rodriguez JA, Griffith KA, Holberg CJ, Martinez FD, Wright AL, Siblings, day-care attendance, and the risk of asthma and wheezing during childhood. N Engl J Med 2000;343: 538-543.

47. Von Ehrenstein OS, Von Mutius E, Illi S, Baumann L, Bohm O, von Kries R, Reduced risk of hay fever and asthma among children of farmers. Clin Exp Allergy 2000;30: 187-193.

48. Wu P, Dupont WD, Griffin MR, et al., Evidence of a causal role of winter virus infection during infancy in early childhood asthma. Am J Respir Crit Care Med 2008;178: 1123-1129.

49. Thomsen SF, van der Sluis S, Stensballe LG, et al., Exploring the association between severe respiratory syncytial virus infection and asthma: a registry-based twin study. Am J Respir Crit Care Med 2009;179: 1091-1097.

50. Silverman M, Grigg J, Mc Kean M, Virus-induced wheeze in young children - A separate disease? In: Johnston S, Papadopoulos N eds. Respiratory infections in allergy and asthma. New York: Marcel Dekker, 2002:427-471.

51. Henderson J, Granell R, Heron J, et al., Associations of wheezing phenotypes in the first 6 years of life with atopy, lung function and airway responsiveness in mid-childhood. Thorax 2008;63: 974-980.

52. Godden DJ, Ross S, Abdalla M, et al., Outcome of wheeze in childhood. Symptoms and pulmonary function 25 years later. Am J Respir Crit Care Med 1994;149: 106-112.

53. Boesen I, Asthmatic bronchitis in children; prognosis for 162 cases, observed 6-11 years. Acta Paediatr 1953;42: 87-96.

54. Global Initiative for Asthma (GINA), Global strategy for asthma management and prevention. 2008 (update): Available from: www.ginasthma.org. Accessed 22 Oct. 2009.

55. Bel EH, Clinical phenotypes of asthma. Curr Opin Pulm Med 2004;10: 44-50.

56. Xu R, Wunsch D, Survey of clustering algorithms. IEEE Trans Neural Netw 2005;16: 645-678.

57. Xu B, Jarvelin MR, Hartikainen AL, Pekkanen J, Maternal age at menarche and atopy among offspring at the age of 31 years. Thorax 2000;55: 691-693.

58. Tuikkala J, Elo LL, Nevalainen OS, Aittokallio T, Missing value imputation improves clustering and interpretation of gene expression microarray data. Bmc Bioinformatics 2008;9: 202.

59. Halkidi M, Batistakis Y, Vazirgiannis M, On clustering validation techniques. J Intell Inf Syst 2001;17: 107-145.

60. Meehl PE, Bootstraps taxometrics - solving the classification problem in psychopathology. Am Psychol 1995;50: 266-275.

61. Ruscio J, Kaczetow W, Differentiating categories and dimensions: evaluating the robustness of taxometric analyses. Multivariate Behav Res 2009;44: 259-280.

62. Ruscio J, Walters GD, Using comparison data to differentiate categorical and dimensional data by examining factor score distributions: resolving the mode problem. Psychol Assessment 2009;21: 578-594.

63. Lubke G, Neale M, Distinguishing between latent classes and continuous factors with categorical outcomes: class invariance of parameters of factor mixture models. Multivariate Behav Res 2008;43: 592-620.

64. Lubke G, Neale MC, Distinguishing between latent classes and continuous factors: resolution by maximum likelihood? Multivariate Behav Res 2006;41: 499-532.

65. Lubke GH, Muthèn B, Moilanen IK, et al., Subtypes versus severity differences in attention-deficit/hyperactivity disorder in the Northern Finnish Birth Cohort. J Am Acad Child Adolesc Psychiatry 2007;46: 1584-1593.

66. Muthen B, Should substance use disorders be considered as categorical or dimensional? Addiction 2006;101 Suppl 1: 6-16.

67. Cox DR, Tests of separate families of hypotheses. Proc 4th Berkeley Symp on Math Statist and Prob 1961;1: 105-123.

68. Cox DR, Further results on tests of separate families of hypotheses. J R Stat Soc Series B Stat Methodol 1962;24: 406-424.

69. Vuong QH, Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 1989;57: 307-333.

70. Pesaran MH, Pesaran B, A simulation approach to the problem of computing coxs statistic for testing nonnested models. J Econometrics 1993;57: 377-392.

71. Rao CR, A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. Qüestiió 1995;19: 23-63.

72. Greenacre MJ, Theory and applications of correspondence analysis. London: Academic Press, 1984.

73. Lazarsfeld PF, Henry NW, Latent structure analysis. Boston: Houghton Mifflin, 1968.

74. Anderson TW, Rubin H, Statistical inference in factor analysis. Proc 4th Berkeley Symp on Math Statist and Prob 1956;1: 111-150.

75. Bartholomew DJ, Factor-analysis for categorical-data. J R Stat Soc Series B Stat Methodol 1980;42: 293-321.

76. Skrondal A, Rabe-Hesket S, Latent variable modelling: A survey. Scand J Statist 2007;34: 712-745.

77. Rabe-Hesketh S, Skrondal A, Pickles A, GLLAMM Manual. UC Berkeley Division of Biostatistics Working Paper Series 2004;Working Paper 160: http://www.bepress.com/ucbbiostat/paper160

78. Muthén B, Latent variable modelling. In: Marcoulides GA, Schumacker RE eds. New developments and techniques in structural equation modeling: Lawrence Erlbaum Associates, 2001:1-33.

79. Muthén B, Beyond SEM: general latent variable modeling. Behaviormetrika 2002;29: 81-117.

80. Muthén BO, Mplus technical appendices. Los Angeles: CA: Muthén & Muthén, 1998-2004.

81. Lubke GH, Muthèn B, Investigating population heterogeneity with factor mixture models. Psychol Methods 2005;10: 21-39.

82. McLachlan G, Peel D, Finite mixture models. New York: John Wiley & Sons, 2000.

83. McLachlan GJ, On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Appl Stat 1987;36: 318-324.

84. Dempster AP, Laird NM, Rubin DB, Maximum likelihood from incomplete data via the EM algorithm. J Roy Statist Soc Ser B 1977;39: 1-38.

85. Schafer JL, Graham JW, Missing data: our view of the state of the art. Psychol Methods 2002;7: 147-177.

86. Hunt L, Jorgensen M, Mixture model clustering using the MULTIMIX program. Aust N Z J Stat 1999;41: 153-171.

87. Hunt L, Jorgensen M, Mixture model clustering for mixed data with missing information. Comput Stat Data Anal 2003;41: 429-440.

88. Khoury MJ, Beaty TH, Cohen BH, Fundamentals of genetic epidemiology. New York: Oxford University Press, 1993.

89. Snider GL, Nosology for our day: its application to chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2003;167: 678-683.

90. Keil T, Kulig M, Simpson A, et al., European birth cohort studies on asthma and atopic diseases: I. Comparison of study designs -- a GALEN initiative. Allergy 2006;61: 221-228.

91. Keil T, Kulig M, Simpson A, et al., European birth cohort studies on asthma and atopic diseases: II. Comparison of outcomes and exposures--a GA2LEN initiative. Allergy 2006;61: 1104-1111.

92. Fens N, Zwinderman AH, van der Schee MP, et al., Exhaled breath profiling enables discrimination of chronic obstructive pulmonary disease and asthma. Am J Respir Crit Care Med 2009;180: 1076-1082.

93. Moeller A, Diefenbacher C, Lehmann A, et al., Exhaled nitric oxide distinguishes between subgroups of preschool children with respiratory symptoms. J Allergy Clin Immunol 2008;121: 705-709.

94. Thornton-Wells TA, Moore JH, Haines JL, Genetics, statistics and human disease: analytical retooling for complexity. Trends Genet 2004;20: 640-647.

95. Ziegler A, Konig IR, Thompson JR, Biostatistical aspects of genome-wide association studies. Biom J 2008;50: 8-28.

96. Brand PL, Baraldi E, Bisgaard H, et al., Definition, assessment and treatment of wheezing disorders in preschool children: an evidence-based approach. Eur Respir J 2008;32: 1096-1110.

Section D: Related Publications

D.1. Cohort profile: The Leicester Respiratory Cohorts (*Int J Epidemiol* 2007;36:977-985)

D.2. Article: Routine vaccination against pertussis and the risk of childhood asthma: A population-based cohort study (*Pediatrics* 2009;123:944-950)

D.3. Correspondence: Timing of routine vaccinations and the risk of childhood asthma (*J Allergy Clin Immunol* 2008; 122:656)

D.4. Editorial: Causal links between RSV infection and asthma – No clear answers to an old question (*Am J Respir Crit Care Med* 2009; 179:1079-80)

D.5. Authors reply: A role for genes and environment in the causal relationship between infant RSV infection and childhood asthma (*Am J Respir Crit Care Med*; in press)

Acknowledgements

I much enjoyed the work on this project and am indebted to many people.

First, I would like to thank my two supervisors Claudia Kuehni and Lutz Dümbgen for their support and competent guidance. They both have invested many hours of supervision. Claudia has put enormous effort into this project, not only during but also before my PhD, setting up 1998 Leicester cohort study. It's a pleasure to work in such highly motivated team. I greatly appreciate Claudia's openness to unconventional approaches and the fruitful application of these ideas would not have been possible without her expertise. Lutz's support was extremely valuable for this project and for my own learning. Having his mathematical and statistical oversight was reassuring and he broadened my view to the more fundamental statistical concepts behind the methods I was applying.

A special thank you goes to Mike Silverman, who together with Claudia initiated this project and who was the architect behind the scene. It was motivating to know that someone who knew so much about the field was backing this work. Mike took much interest in this work and invested many fruitful hours discussing progress and giving his invaluable advice.

I thank John Thompson, the co-referee of this thesis, for his guidance. Though the times we met were fewer, his input was very helpful. John prompted me to find a focus for the 2nd phase of my PhD and his advice came at a moment when it was much needed.

Also, I want to thank Christoph Minder, former senior statistician at ISPM, who was significantly involved in some of this work, particularly in an early phase of my PhD, giving me much advice and helping me find a methodological orientation for my work.

I particularly want to thank the other members of our asthma team here in Bern, Marie, Cris and Nora. Marie who has been my constant companion and office mate throughout the last years has seen me through my ups and downs. She has also contributed a lot to this work, doing most of the data management, participating in many long discussions and providing technical support. Cris and Nora have both helped in proofreading parts of this thesis. Thanks also to Cris for sacrificing some of your time for this during Christmas season.

Similarly I want to thank the members of the Leicester team, particularly Caroline Beardsmore, Teresa McNally, Manjith Narayanan, Michelle Moore, Ketna Parmar, Siân

Williams. It is a great pleasure working with all of you and spending time together in crammed offices, at curry dinners and in deep snow up in the Alps. I also thank Kathryn Staley who has always been supportive, when I was in Leicester or when she came to Bern - I count her in on the Leicester team. Kathryn has also proofread parts of the thesis.

I also want to thank my friends and colleagues here at the ISPM Bern, the fellow-PhD students for their friendship and the fun times we had, the senior staff, the librarians and secretaries for their continual support, in particular also Martin Brinkhof with whom I spent many hours of work but also of inspiring discussions, all the people from the cancer registry (without naming them all). Equally I want to thank my friends and colleagues from the paediatric respiratory group at the Inselspital for their friendship and support, in particular also Urs Frey, Philip Latzin, Cindy Thamrin, Oliver Fuchs for their thoughts and comments.

I want to thank my parents, my sisters and brothers, and my friends for being so supportive of me during this time.

Curriculum Vitae

Personal information

Name	Ben Daniel Spycher
Address	Finkenhubelweg 11 CH- 3012 Bern, Switzerland
Phone	+41 (031) 631 35 07
Fax	+41 (031) 631 35 20
e-mail	bspycher@ispm.unibe.ch
Date of birth	May 22, 1971
Place of birth	Bern, Switzerland
Nationality	Swiss
Languages	German, English, French (all spoken and written)

Professional experience

2006-present	PhD student, Institute of Social and Preventive Medicine (ISPM), University of Bern, Switzerland
2005-2006	Research Fellow, Institute of Social and Preventive Medicine (ISPM), University of Bern, Switzerland
1999-2001	Associate Professional Officer, International Livestock Research Institute (ILRI), Niamey, Niger, West Africa
1998-99	Research Fellow, Institute of Agricultural Economics, ETH-Zürich, Switzerland

Education

1/10/2005	BSc in Statistics, Freie Universität Berlin, Germany
5/5/1998	MSc in agricultural economics (Diplom als Ingenieur-Agronom), Swiss Federal Institute of Technology Zürich (ETH-Zürich), Switzerland

Courses attended

2009	Competing Risks (1 day). ISPM, Bern
	The Complex Genome (2 days). Department of Social Medicine, Bristol
2008	Practical Bayesian Models for the Health Sciences (1 day). ISPM, Bern
	Genetic epidemiology of complex diseases: Principles and methods (3 days). IUMSP, Lausanne

2007	Causal Inference from Observational Data (3 days). ISPM, Bern
	Workshop on Latent Variable Modelling with Mplus (3 days). Department of Statistics, Florence
	Methods for Dealing with Missing Data (3 days). Department of Social Medicine, Bristol
	Advanced Methods in Epidemiology: Analysis of Clustered Data and Multilevel Modelling (5 days). ISPM, Bern
2006	Writing a Journal Article and Getting it Published (3 days). ISPM, Bern
	Advanced Methods in Epidemiology: Applied Regression Modelling (5 days). ISPM, Bern
2005	Advanced Methods in Epidemiology: Meta-Analysis and Systematic Reviews (5 days). ISPM, Bern

Tutoring / Teaching

2009	Faculty of Medicine, University of Bern:
	MPH course: Introduction to Stata10 (2.5 hours)
	Book club: Modelling Infectious Diseases, Stochastic dynamics (1 hour)
2007	Faculty of Medicine, University of Bern:
	MPH course: Analysis of Clustered Data and Multilevel Modelling (1 hour)
	Graduate School for Cellular and Biomedical Sciences, University of Bern:
	Book club: Cancer epidemiology, Measures of disease occurrence (1 hour)
	Book club: Cancer epidemiology, Survival analysis (2 hours)
2006	Faculty of Medicine, University of Bern:
	MPH course: Applied Regression Modelling (10 hours)

Reviewing manuscripts

European Respiratory Journal

Clinical and Experimental Allergy

Pediatric Pulmonology

Allergy

List of Publications

Publications in peer-reviewed journals

Original articles

2009 1. **Spycher BD**, Silverman M, Barben B, Eber E, Guinand S, Levy ML, Pao C, van Aalderen WM, van Schayck OCP, Kuehni CE. A disease model for wheezing disorders in preschool children based on clinicians' perceptions. *PLoS ONE 2009; 4:e8533.*

2. **Spycher BD**, Minder CE, Kuehni CE. Multivariate modelling of responses to conditional items: new possibilities for latent class analysis. *Stat Med 2009; 28:1927-39.*

3. **Spycher BD**, Silverman M, Zwahlen M, Brooke AM, Kuehni CE. Routine vaccination against pertussis and the risk of childhood asthma: a population-based cohort study. *Pediatrics 2009; 123:944-50.*

2008 4. **Spycher BD**, Silverman M, Brooke AM, Minder CE, Kuehni CE. Distinguishing phenotypes of childhood wheeze and cough: a novel approach with prognostic relevance. *Eur Respir J 2008; 31:974-81.*

2007 5. Staley KG, Strippoli MPF, **Spycher BD**, Stower C, Silverman M, Kuehni CE. Mannan-binding lectin in young children with asthma differs by level of severity. *J Allergy Clin Immunol 2007; 119:503-5.*

6. Kuehni CE, Brooke AM, Strippoli MPF, **Spycher BD**, Davis A, Silverman M. Cohort profile: The Leicester Respiratory Cohorts. *Int J Epidemiol 2007; 36:977-85.*

Editorials / Correspondence

2009 7. Kuehni CE, **Spycher BD**, Silverman M. Causal links between RSV infection and asthma – No clear answers to an old question. *Am J Respir Crit Care Med 2009; 179:1079-80*.

8. Kuehni CE, **Spycher BD**, Silverman M. role for genes and environment in the causal relationship between infant RSV infection and childhood asthma. (Correspondence: Reply to Wu P, Dupont WD, Griffin MR, Hartert TV. Am J Respir Crit Care Med; in press) *Am J Respir Crit Care Med; in press.*

2008 9. **Spycher BD**, Silverman M, Kuehni CE. Timing of routine vaccinations and the risk of childhood asthma. (Correspondence: Reply to McDonald et al. J Allergy Clin Immunol 2008; 121:626-31) *J Allergy Clin Immunol 2008; 122:656.*

Reviews

2009 10. **Spycher BD**, Silverman M, Kuehni CE. Phenotypes of childhood asthma: are they real? *Clin Exp Allergy; submitted.*

Other publications

2006
1. Snitzman J, **Spycher BD**, Frey U, Wildhaber JH. Direct maximum expiratory flow modelling from lung function testing of pediatric patients. *In: Proceedings of the 5th World Congress of Biomechanics, 29 July-4 August 2006, Munich, Germany, edited by D. Liepsch: Medimond Inter. Proc., pp. 319-324, 2006.*
2. Williams TO, Okike I, **Spycher BD**. A Hedonic Analysis of Cattle Prices in the Central Corridor of West Africa: Implications for Production and Marketing Decisions. *In: International Association of Agricultural Economist Conference, August 12-18, 2006, Queensland, Australia, http://purl.umn.edu/25423.*

Abstracts in peer-reviewed journals

2009
1. Strippoli MPF, **Spycher BD**, Silverman M, Beardsmore CS, Kuehni CE. Paracetamol use and the risk of wheeze: causation or bias? *Eur Respir J 2009;*
2. Kuehni CE, Strippoli MPF, **Spycher BD**, Silverman M, Beardsmore CS. Early daycare and the risk of wheeze from birth through 10 years of age. *Eur Respir J 2009;*
3. **Spycher BD**, Silverman M, Strippoli MPF, Kuehni CE. Childhood wheeze: one or several diseases? *Eur Respir J 2009;*
4. Carlsen-Lodrup K, Roll S, …, **Spycher BD**, Kuehni CE, … Keil T. Pet exposure in infancy, asthma at school age? A meta-analysis initiated by GA(2)LEN. *Allergy 2009; 64(Suppl):22.*
5. Keil T, Roll S, …, **Spycher BD**, Kuehni CE, … Carlsen-Lodrup K. Pet exposure in infancy, allergis rhinitis at school age? A meta-analysis initiated by GA(2)LEN. *Allergy 2009; 64(Suppl):23.*

2008
6. **Spycher BD**, Strippoli MPF, Silverman M, Kuehni CE. Predicting persistence of childhood wheeze using a symptom based severity score. *Eur Respir J 2008; 32(Suppl):562-3s.*
7. Strippoli MPF, **Spycher BD**, Silverman M, Kuehni CE. Breastfeeding and the risk of childhood asthma: a population-based cohort study. *Eur Respir J 2008; 32(Suppl):772s.*
8. Kuehni C, Strippoli MPF, **Spycher BD**, Silverman M. Prevalence and characteristics of wheezing disorders in the community from age 1 to 9 years. *Eur Respir J 2008; 32(Suppl):562s.*

2007
9. Kuehni CE, Strippoli MPF, **Spycher BD**, McNally T, Silverman M. Predictors of exhaled nitric oxide (FENO) in a large population-based sample of schoolchildren. *Eur Respir J 2007; 30(Suppl):724s.*
10. **Spycher BD**, Silverman M, Strippoli MPF, Minder C, Brooke AM, Kuehni CE. Non-specific chronic cough in children: a novel approach to phenotype identification. *Eur Respir J 2007; 30(Suppl):398s.*
11. Strippoli MPF, Silverman M, **Spycher BD**, McNally T, Kuehni CE. Ethnic differences in atopy and exhaled nitric oxide (FENO) in schoolchildren with wheeze. *Eur Respir J 2007; 30(Suppl):684s*.
12. Kuehni CE, Silverman M, Strippoli MPF, Brooke AM, **Spycher BD**. Non-specific chronic cough in children: a novel method for identification of clinical phenotypes. *Swiss Med Wkly 2007; 137(Suppl):4s.*
13. Strippoli MPF, Silverman M, **Spycher BD**, Brooke AM, Kuehni CE. Wheezing phenotypes in childhood: Physiological characteristics of viral compared to multiple trigger wheezers. *Swiss Med Wkly 2007; 137(Suppl):8s.*

Declaration of Originality

Last name, first name: **Spycher, Ben**

Matriculation number: **92-905-314**

I hereby declare that this thesis represents my original work and that I have used no other sources except as noted by citations.

All data, tables, figures and text citations which have been reproduced from any other source, including the internet, have been explicitly acknowledged as such.

I am aware that in case of non-compliance, the Senate is entitled to divest me of the doctorate degree awarded to me on the basis of the present thesis, in accordance with the "Statut der Universität Bern (Universitätsstatut; UniSt)", Art. 20, of 17 December 1997.

Place, date Signature

…… ……

Appendices

i. Supplementary material: Distinguishing phenotypes of childhood wheeze and cough using latent class analysis (*Eur Resp J* 2008; 31:974-981)

ii. Abstract: Predicting persistence of childhood wheeze using a symptom based severity score (*Eur Respir J* 2008; 32:562s-563s)

iii. Abstract: Childhood wheeze: one or several diseases? (*Eur Respir J* 2008; 34:756s)

iv. Example of multiple correspondence analysis

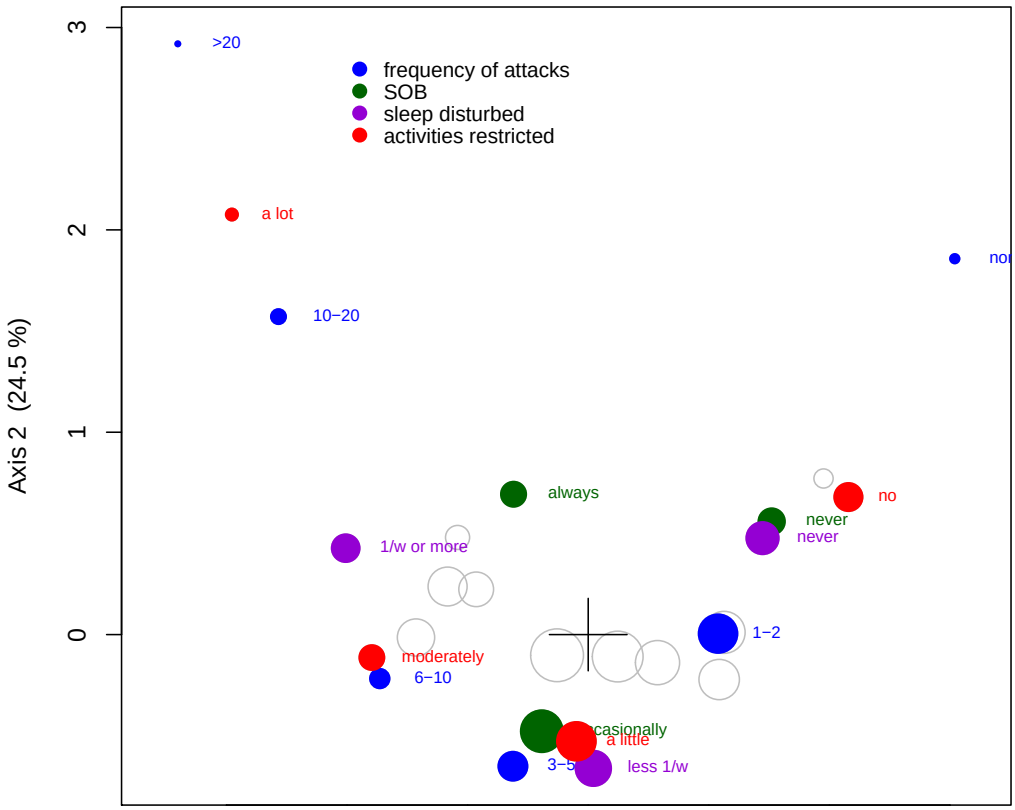
Figure 5: Example of MCA using data on symptoms of wheeze

The Figures show relative positions of response categories related to severity of wheeze (A) and triggers of wheeze (B) from a single MCA analysis of symptoms listed below in 241 children aged 4 years from the 1998(a) Leicester cohort. The U-shape pattern is the so-called "horseshoe-effect" (ch. 8.3 in [72]) indicating the presence of an underlying gradient. The sizes of the circles are proportional to the prevalences of the corresponding response categories. Missing values have been imputed by iteration of correspondence analysis using the procedure proposed in ch. 8.5 in [72].

Variables:

- *Frequency of attacks*: During the past 12 months, how many attacks of wheezing has he/she had? None ; 1-2 ; 3-5 ; 6-10 ; 10-20; more than 20
- *Shortness of breath (SOB)*: Do these attacks cause him/her to be short of breath? yes, always; yes, occasionally; no, never
- *Sleep disturbed*: In the last 12 months, how often, on average, has your child's sleep been disturbed due to wheezing? Never woken with wheezing; less than one night per week; one or more nights per week
- *Activities disturbed*: In the last 12 months, how much did wheezing interfere with your child's daily activities? not at all; a little; a moderate amount; a lot
- *Triggers*: Do these attacks occur: (answer all please)
  - when he/she is running or playing? yes ; no
  - with drinking or eating? yes ; no
  - when he/she is near animals, dust or grass? yes ; no
- *Wheeze with cold*: In the last 12 months, has your child had wheezing or whistling in the chest during or soon after a cold or flu? yes; no
- *Wheeze without cold*: In the last 12 months, has your child had wheezing or whistling in the chest even without having a cold or flu? yes; no

A) Indicators of severity of wheeze



B) Triggers of wheeze